

A Conjugate Property between Loss Functions and Uncertainty Sets in Classification Problems

Takafumi Kanamori
Nagoya University
kanamori@is.nagoya-u.ac.jp

Akiko Takeda
Keio University
takeda@ae.keio.ac.jp

Taiji Suzuki
The University of Tokyo
t-suzuki@mist.i.u-tokyo.ac.jp

Abstract

In binary classification problems, mainly two approaches have been proposed; one is loss function approach and the other is uncertainty set approach. The loss function approach is applied to major learning algorithms such as support vector machine (SVM) and boosting methods. The loss function represents the penalty of the decision function on the training samples. In the learning algorithm, the empirical mean of the loss function is minimized to obtain the classifier. Against a backdrop of the development of mathematical programming, nowadays learning algorithms based on loss functions are widely applied to real-world data analysis. In addition, statistical properties of such learning algorithms are well-understood based on a lots of theoretical works. On the other hand, the learning method using the so-called uncertainty set is used in hard-margin SVM, mini-max probability machine (MPM) and maximum margin MPM. In the learning algorithm, firstly, the uncertainty set is defined for each binary label based on the training samples. Then, the best separating hyperplane between the two uncertainty sets is employed as the decision function. This is regarded as an extension of the maximum-margin approach. The uncertainty set approach has been studied as an application of robust optimization in the field of mathematical programming. The statistical properties of learning algorithms with uncertainty sets have not been intensively studied. In this paper, we consider the relation between the above two approaches. We point out that the uncertainty set is described by using the level set of the conjugate of the loss function. Based on such relation, we study statistical properties of learning algorithms using uncertainty sets.

1 Introduction

In classification problems, the goal is to predict output labels for given input vectors. For this purpose, a decision function defined on the input space is estimated from training samples. The output value of the decision function is used for the label prediction. In binary classification problems, the label is predicted by the sign of the decision function.

Many learning algorithms use loss functions to measure the penalty of misclassifications. The decision function minimizing the empirical mean of the loss function over training samples is employed as the estimator [8, 24, 12, 14]. For example, hinge loss, exponential loss and logistic loss are used for support vector machine (SVM), Adaboost and logistic regression, respectively.

Especially in the binary classification tasks, statistical properties of learning algorithms based on loss functions are well-understood due to intensive recent works. See [2, 26, 25, 22, 30, 29] for details.

As another approach, the maximum-margin criterion is also applied for the statistical learning. Under the maximum-margin criterion, the best separating hyperplane between the two output labels is employed as the decision function. In hard-margin SVM [29], a convex-hull of input vectors for each binary label is defined, and the maximum-margin between the two convex-hulls is considered. For the non-separable case, ν -SVM provides a similar picture [24, 5]. In ν -SVM, the so-called reduced convex-hull which is a subset of the original convex-hull is used for the learning. A reduced convex-hull is defined for each label, and the best separating hyperplane between the two reduced convex-hulls is employed as the decision function. Not only polyhedral sets such as the convex-hull of finite input points but also ellipsoidal sets are applied for classification problems [15, 18]. In this paper, the set used in the maximum-margin criterion is referred to as *uncertainty set*. This term is borrowed from robust optimization in mathematical programming [4].

There are some works in which the statistical properties of the learning based on the uncertainty set are studied. For example, [15] proposed minimax probability machine (MPM) using the ellipsoidal uncertainty sets, and studied statistical properties under the worst-case setting. In the statistical learning using uncertainty set, the main concern is to develop optimization algorithms under the maximum margin criterion [17]. So far, statistical properties of the learning algorithm using uncertainty sets have not been intensively studied compared to the learning using loss functions.

The main purpose of this paper is to study the learning algorithm using the uncertainty set. We focus on the relation between the loss function and the uncertainty set. We show that the uncertainty set is described by using the conjugate function of the loss function. For given uncertainty set, we construct the corresponding loss function. We study the statistical properties of the learning algorithm using the uncertainty set by applying theoretical results on the loss function approach. Then, we establish the statistical consistency of learning algorithms using the uncertainty set. We point out that in general the maximum margin criterion for a fixed uncertainty set does not provide accurate decision functions. We need to introduce a parametrized uncertainty set by the one-dimensional parameter which specifies the size of the uncertainty set. We show that a modified maximum margin criterion with the parametrized uncertainty set recovers the statistical consistency.

The paper is organized as follows. In Section 2, we introduce the existing method based on the uncertainty set. In Section 3, we investigate the relation between loss functions and uncertainty sets. Section 4 is devoted to illustrate a way of revising the uncertainty set to recover nice statistical properties. In Section 5, we present a kernel-based learning algorithm with uncertainty sets. In Section 6, we prove that the proposed algorithm has the statistical consistency. Numerical experiments are shown in Section 7. We conclude in section 8. Some proofs are shown in Appendix.

We summarize some notations to be used throughout the paper. The indicator function is denoted as $\mathbb{I}[A]$, i.e., $\mathbb{I}[A]$ equals 1 if A is true, and 0 otherwise. The column vector \mathbf{x} in the Euclidean space is described in bold face. The transposition of \mathbf{x} is denoted as \mathbf{x}^T . The Euclidean norm of the vector \mathbf{x} is expressed as $\|\mathbf{x}\|$. For a set S in a linear space, the convex-hull of S is denoted as $\text{conv}S$ or $\text{conv}(S)$. The number of elements in the set S is denoted as $|S|$. The expectation of the random variable Z w.r.t. the probability distribution P is described as

$\mathbb{E}_P[Z]$. We will drop the subscript P as $\mathbb{E}[Z]$, when it is clear from the context. The set of all measurable functions on the set \mathcal{X} is denoted by $L_0(\mathcal{X})$ or L_0 for short. The supremum norm of $f \in L_0$ is denoted as $\|f\|_\infty$. For the reproducing kernel Hilbert space \mathcal{H} , $\|f\|_{\mathcal{H}}$ is the norm of $f \in \mathcal{H}$ defined from the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ on \mathcal{H} .

2 Preliminaries

We define \mathcal{X} as the input space and $\{+1, -1\}$ as the set of binary labels. Suppose that the training samples $(x_1, y_1), \dots, (x_m, y_m) \in \mathcal{X} \times \{+1, -1\}$ are drawn i.i.d. according to a probability distribution P on $\mathcal{X} \times \{+1, -1\}$. The goal is to estimate a decision function $f : \mathcal{X} \rightarrow \mathbb{R}$ from a set of functions \mathcal{F} , such that the sign of $f(x)$ provides an accurate prediction of the unknown binary label associated with the input x under the probability distribution P . In other word, for the estimated decision function f , the probability of $\text{sign}(f(x)) \neq y$ is expected to be as small as possible. In this article, the composite function of the sign function and the decision function, $\text{sign}(f(x))$, is referred to as classifier.

2.1 Learning with loss functions

In binary classification problems, the prediction accuracy of the decision function f is measured by the 0-1 loss $\mathbb{I}[yf(x) \leq 0]$ which equals 1 when the sign of $f(x)$ is different from y and 0 otherwise. The average prediction performance of the decision function f is evaluated by the expected 0-1 loss, i.e.,

$$\mathcal{E}(f) = \mathbb{E}[\mathbb{I}[yf(x) \leq 0]]. \quad (1)$$

The Bayes risk \mathcal{E}^* is defined as the minimum value of the expected 0-1 loss over all the measurable functions on \mathcal{X} ,

$$\mathcal{E}^* = \inf\{\mathcal{E}(f) : f \in L_0\}. \quad (2)$$

Bayes risk is the lowest achievable error rate under the probability P . Given the set of training samples, $T = \{(x_1, y_1), \dots, (x_m, y_m)\}$, the empirical 0-1 loss is denoted by

$$\hat{\mathcal{E}}_T(f) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}[y_i f(x_i) \leq 0]. \quad (3)$$

The subscript T in $\hat{\mathcal{E}}_T(f)$ is dropped if it is clear from the context.

In general, minimization of $\hat{\mathcal{E}}_T(f)$ is considered as a hard problem [1]. The main difficulty is considered to come from non-convexity of the 0-1 loss $\mathbb{I}[yf(x) \leq 0]$ as the function of f . Hence, many learning algorithms use a surrogate loss of the 0-1 loss in order to make the computation tractable. For example, SVM uses the hinge loss, $\max\{1 - yf(x), 0\}$, and Adaboost uses the exponential loss, $\exp\{-yf(x)\}$. Both the hinge loss and the exponential loss are convex in f , and they provide an upper bound of the 0-1 loss. Thus, the minimizer under the surrogate loss is also expected to minimize the 0-1 loss. The quantitative relation between the 0-1 loss and the surrogate loss was studied by [2].

To avoid overfitting of the estimated decision function to training samples, the regularization is considered. By adding the regularization term such as the squared norm of the decision function to the empirical surrogate loss, the complexity of the estimated classifier is restricted. The

balance between the regularization term and the surrogate loss is adjusted by the regularization parameter [11, 26]. Then, the deviation of the empirical 0-1 loss and the expected 0-1 loss is controlled by the regularization. When both the regularization term and the surrogate loss are convex, the computational tractability of the statistical learning is retained.

2.2 Learning with uncertainty sets

Besides statistical learning using loss functions, there is another approach to the classification problems, i.e., statistical learning based on the so-called *uncertainty set*. We briefly introduce the basic idea of the uncertainty set. We assume that \mathcal{X} is a subset of Euclidean space.

In robust optimization problems [4], the uncertainty set describes uncertainties or ambiguities included in optimization problems. The parameter in the optimization problem may not be precisely determined. Instead of the precise information, we have an uncertainty set which probably includes the parameter in the optimization problem. The worst-case setting is employed to solve the robust optimization problem with the uncertainty set.

The statistical learning with uncertainty set is considered as an application of the robust optimization to classification problems. In classification problems, the uncertainty set is designed such that most training samples are included in the uncertainty set with high probability. We prepare an uncertainty set for each binary label. For example, \mathcal{U}_p and \mathcal{U}_n are the confidence regions such that the conditional probabilities, $P(\mathbf{x} \in \mathcal{U}_p | y = +1)$ and $P(\mathbf{x} \in \mathcal{U}_n | y = -1)$, are equal to 0.95. As the other example, the uncertainty set \mathcal{U}_p (resp. \mathcal{U}_n) consists of the convex-hull of input vectors in training samples having the positive (resp. negative) label. The convex-hull of data points is used in hard margin SVM [5]. The ellipsoidal uncertainty set is also used for the robust classification under the worst-case setting [15, 18].

Based on the uncertainty set, we estimate the linear decision function $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$. Here, we consider the *minimum distance problem*

$$\min_{\mathbf{x}_p, \mathbf{x}_n} \|\mathbf{x}_p - \mathbf{x}_n\| \quad \text{subject to} \quad \mathbf{x}_p \in \mathcal{U}_p, \mathbf{x}_n \in \mathcal{U}_n. \quad (4)$$

Let \mathbf{x}_p^* and \mathbf{x}_n^* be optimal solutions of (4). Then, the normal vector of the decision function, \mathbf{w} , is estimated by $c(\mathbf{x}_p^* - \mathbf{x}_n^*)$, where c is a positive real number. Figure 1 illustrates the estimated decision boundary. When both \mathcal{U}_p and \mathcal{U}_n are compact subsets satisfying $\mathcal{U}_p \cap \mathcal{U}_n = \emptyset$, the estimated normal vector cannot be the null vector. The minimum distance problem appears in the hard margin SVM [29, 5], ν -SVM [24, 9] and the learning algorithms proposed by [18, 17]. In Section 3.1, we briefly introduce the relation between ν -SVM and the minimum distance problem. In minimax probability machine (MPM) proposed by [15], the other criterion is applied to estimate the linear decision function, though the ellipsoidal uncertainty set plays an important role also in their algorithm.

The minimum distance problem is equivalent with the maximum margin principle [29, 5]. When the bias term b in the linear decision function is estimated such that the decision boundary bisects the line segment connecting \mathbf{x}_p^* and \mathbf{x}_n^* , the estimated decision boundary achieves the maximum margin between the uncertainty sets, $\mathcal{U}_p, \mathcal{U}_n$. According to [28], we explain how the maximum margin is connected with the minimum distance. Suppose that \mathcal{U}_p and \mathcal{U}_n are convex subsets and that $\mathcal{U}_p \cap \mathcal{U}_n = \emptyset$ holds. Then, the margin of two uncertainty sets along the direction

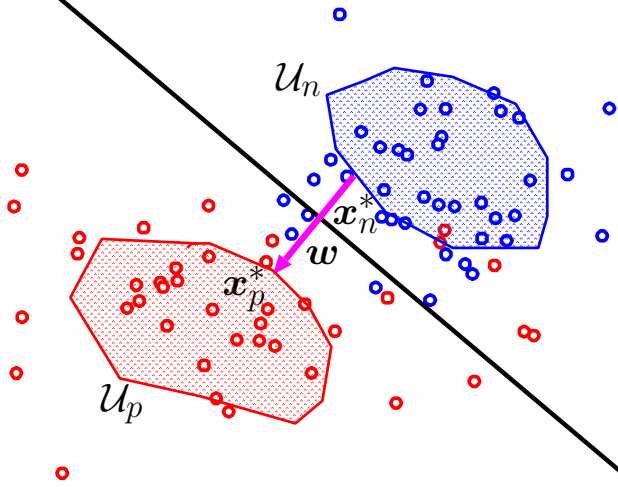


Figure 1: The estimated decision boundary based on the minimum distance problem with the uncertainty sets \mathcal{U}_p and \mathcal{U}_n .

of \mathbf{w} is given as

$$\min \left\{ \frac{\mathbf{w}^T \mathbf{x}_p - \mathbf{w}^T \mathbf{x}_n}{\|\mathbf{w}\|} : \mathbf{x}_p \in \mathcal{U}_p, \mathbf{x}_n \in \mathcal{U}_n \right\}.$$

The maximum margin criterion is described as

$$\max_{\mathbf{w} \neq \mathbf{0}} \min \left\{ \frac{\mathbf{w}^T \mathbf{x}_p - \mathbf{w}^T \mathbf{x}_n}{\|\mathbf{w}\|} : \mathbf{x}_p \in \mathcal{U}_p, \mathbf{x}_n \in \mathcal{U}_n \right\} = \min \{\|\mathbf{x}_p - \mathbf{x}_n\| : \mathbf{x}_p \in \mathcal{U}_p, \mathbf{x}_n \in \mathcal{U}_n\}.$$

The equality above follows from the minimum norm duality [16].

3 Relation between Loss Functions and Uncertainty Sets

We study the relation between loss functions and uncertainty sets. First, we introduce the relation in ν -SVM according to [9] and [5]. Then, we present an extension of ν -SVM to investigate a generalized relation between loss functions and uncertainty sets.

3.1 Uncertainty Set in ν -SVM

Suppose that the input space \mathcal{X} is a subset of Euclidean space \mathbb{R}^d . We consider the linear decision function, $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$, where the normal vector $\mathbf{w} \in \mathbb{R}^d$ and the bias term $b \in \mathbb{R}$ are to be estimated based on observed training samples. By applying the kernel trick [6, 23], we obtain rich statistical models for the decision function, while keeping the computational tractability.

In ν -SVM, the classifier is estimated as the optimal solution of

$$\min_{\mathbf{w}, b, \rho} \frac{1}{2} \|\mathbf{w}\|^2 - \nu \rho + \frac{1}{m} \sum_{i=1}^m \max\{\rho - y_i(\mathbf{w}^T \mathbf{x}_i + b), 0\}, \quad \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}, \rho \in \mathbb{R}, \quad (5)$$

where $\nu \in (0, 1)$ is a prespecified constant which has the role of the regularization parameter. As [24] pointed out, the parameter ν controls the margin errors and number of support vectors. In ν -SVM, a variant of the hinge loss, $\max\{\rho - y_i(\mathbf{w}^T \mathbf{x}_i + b), 0\}$, is used as the surrogate loss. In the original formulation of ν -SVM, the non-negativity constraint, $\rho \geq 0$, is introduced. As shown by [9], we can confirm that the non-negativity constraint is redundant. Indeed, for an optimal solution $\hat{\mathbf{w}}, \hat{b}, \hat{\rho}$, we have

$$-\nu\hat{\rho} \leq \frac{1}{2}\|\hat{\mathbf{w}}\|^2 - \nu\hat{\rho} + \frac{1}{m} \sum_{i=1}^m \max\{\hat{\rho} - y_i(\hat{\mathbf{w}}^T \mathbf{x}_i + \hat{b}), 0\} \leq 0,$$

where the last inequality comes from the fact that the parameter, $\mathbf{w} = \mathbf{0}, b = 0, \rho = 0$, is a feasible solution of (5). As a result, we have $\hat{\rho} \geq 0$ for $\nu > 0$.

We briefly show that the dual problem of (5) yields the minimum distance problem in which the reduced convex-hulls of training samples are used as uncertainty sets. See [5] for details. The problem (5) is equivalent with

$$\begin{aligned} \min_{\mathbf{w}, b, \rho, \xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{m} \sum_{i=1}^m \xi_i, \\ \text{subject to} \quad & \xi_i \geq 0, \quad \xi_i \geq \rho - y_i(\mathbf{w}^T \mathbf{x}_i + b), \quad i = 1, \dots, m. \end{aligned}$$

Then, the Lagrangian function is defined as

$$L(\mathbf{w}, b, \rho, \xi, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{m} \sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i(\rho - y_i(\mathbf{w}^T \mathbf{x}_i + b) - \xi_i) - \sum_{i=1}^m \beta_i \xi_i,$$

where $\alpha_i, \beta_i, i = 1, \dots, m$ are non-negative Lagrange multipliers. For the observed training samples, we define M_p and M_n as the set of sample indices for each label, i.e.,

$$M_p = \{i \mid y_i = +1\}, \quad M_n = \{i \mid y_i = -1\}. \quad (6)$$

By applying min-max theorem, we have

$$\begin{aligned} & \inf_{\mathbf{w}, b, \rho, \xi} \sup_{\alpha \geq 0, \beta \geq 0} L(\mathbf{w}, b, \rho, \xi, \alpha, \beta) \\ &= \sup_{\alpha \geq 0, \beta \geq 0} \inf_{\mathbf{w}, b, \rho, \xi} L(\mathbf{w}, b, \rho, \xi, \alpha, \beta) \\ &= \sup \left\{ -\frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 : \sum_{i=1}^m \alpha_i = \nu, \sum_{i=1}^m \alpha_i y_i = 0, 0 \leq \alpha_i \leq \frac{1}{m} \right\} \\ &= -\frac{\nu^2}{8} \inf \left\{ \left\| \sum_{i \in M_p} \gamma_i \mathbf{x}_i - \sum_{j \in M_n} \gamma_j \mathbf{x}_j \right\|^2 : \sum_{i \in M_p} \gamma_i = \sum_{i \in M_n} \gamma_i = 1, 0 \leq \gamma_i \leq \frac{2}{m\nu}, i = 1, \dots, m \right\}, \end{aligned} \quad (7)$$

where the last equality is obtained by changing the variable from α_i to $\gamma_i = 2\alpha_i/\nu$. For the positive (resp. negative) label, we introduce the uncertainty set \mathcal{U}_p (reps. \mathcal{U}_n) defined by the reduced convex-hull, i.e.,

$$o \in \{p, n\}, \quad \mathcal{U}_o = \left\{ \sum_{i \in M_o} \gamma_i \mathbf{x}_i : \sum_{i \in M_o} \gamma_i = 1, 0 \leq \gamma_i \leq \frac{2}{m\nu}, i \in M_o \right\}.$$

When the upper limit of γ_i is less than one, the reduced convex-hull is a subset of the convex-hull of training samples. We find that solving the problem (7) is identical to solving the minimum distance problem under the uncertainty set of the reduced convex-hulls,

$$\inf_{\mathbf{x}_p, \mathbf{x}_n} \|\mathbf{x}_p - \mathbf{x}_n\| \quad \text{subject to} \quad \mathbf{x}_p \in \mathcal{U}_p, \mathbf{x}_n \in \mathcal{U}_n.$$

The representation based on the minimum distance problem provides an intuitive understanding of the learning algorithm.

3.2 Uncertainty Set Associated with Loss Function

We consider general loss functions, and study the relation between the loss function and the corresponding uncertainty set. Again, the decision function is defined as $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$ on \mathbb{R}^d . Let $\ell : \mathbb{R} \rightarrow \mathbb{R}$ be a convex and non-decreasing function. For the training samples, $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$, we propose a learning method in which the decision function is estimated by solving

$$\inf_{\mathbf{w}, b, \rho} -2\rho + \frac{1}{m} \sum_{i=1}^m \ell(\rho - y_i(\mathbf{w}^T \mathbf{x}_i + b)) \quad \text{subject to} \quad \|\mathbf{w}\|^2 \leq \lambda^2, b \in \mathbb{R}, \rho \in \mathbb{R}. \quad (8)$$

The regularization effect is introduced by the constraint $\|\mathbf{w}\|^2 \leq \lambda^2$, where λ is the regularization parameter which may depend on the sample size.

The statistical learning using (8) is regarded as an extension of ν -SVM. To see this, we define $\ell(z) = \max\{2z/\nu, 0\}$. Let $\hat{\mathbf{w}}, \hat{b}, \hat{\rho}$ be an optimal solution of (5) for a fixed $\nu \in (0, 1)$. By comparing the optimality conditions of (5) and (8), we can confirm that the problem (8) with $\lambda = \|\hat{\mathbf{w}}\|$ has the same optimal solution as ν -SVM.

In the similar way as ν -SVM, we derive the uncertainty set associated with the loss function ℓ in (8). We introduce the slack variables $\xi_i, i = 1, \dots, m$ satisfying the inequalities $\xi_i \geq \rho - y_i(\mathbf{w}^T \mathbf{x}_i + b)$, $i = 1, \dots, m$. Then, the Lagrangian function of (8) is given as

$$L(\mathbf{w}, b, \rho, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mu) = -2\rho + \frac{1}{m} \sum_{i=1}^m \ell(\xi_i) + \sum_{i=1}^m \alpha_i (\rho - y_i(\mathbf{w}^T \mathbf{x}_i + b) - \xi_i) + \mu(\|\mathbf{w}\|^2 - \lambda^2),$$

where $\alpha_1, \dots, \alpha_m$ and μ are the non-negative Lagrange multipliers. The optimality conditions,

$$\frac{\partial L}{\partial \rho} = 0, \quad \text{and} \quad \frac{\partial L}{\partial b} = 0$$

and the non-negativity of α_i lead to the constraint on Lagrange multipliers,

$$\sum_{i \in M_p} \alpha_i = \sum_{i \in M_n} \alpha_i = 1, \quad \alpha_i \geq 0.$$

We define the conjugate function of $\ell(z)$ as

$$\ell^*(x) = \sup_{z \in \mathbb{R}} \{xz - \ell(z)\}.$$

Then, by applying min-max theorem, we have

$$\begin{aligned}
& \inf_{\mathbf{w}, b, \rho, \xi} \sup_{\alpha \geq 0, \mu \geq 0} L(\mathbf{w}, b, \rho, \xi, \alpha, \mu) \\
&= \sup_{\alpha \geq 0, \mu \geq 0} \inf_{\mathbf{w}, b, \rho, \xi} L(\mathbf{w}, b, \rho, \xi, \alpha, \mu) \\
&= \sup_{\alpha, \mu \geq 0} \inf_{\mathbf{w}, \xi} \left\{ -\frac{1}{m} \sum_{i=1}^m (m\alpha_i \xi_i - \ell(\xi_i)) - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^T \mathbf{w} + \mu (\|\mathbf{w}\|^2 - \lambda^2) \right. \\
&\quad \left. : \sum_{i \in M_p} \alpha_i = \sum_{i \in M_n} \alpha_i = 1, \alpha_i \geq 0 \right\} \\
&= - \inf_{\alpha, \mu \geq 0} \left\{ \frac{1}{m} \sum_{i=1}^m \ell^*(m\alpha_i) + \frac{1}{4\mu} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 + \mu \lambda^2 : \sum_{i \in M_p} \alpha_i = \sum_{i \in M_n} \alpha_i = 1, \alpha_i \geq 0 \right\} \\
&= - \inf_{\alpha} \left\{ \frac{1}{m} \sum_{i=1}^m \ell^*(m\alpha_i) + \lambda \left\| \sum_{i \in M_p} \alpha_i \mathbf{x}_i - \sum_{i \in M_n} \alpha_i \mathbf{x}_i \right\| : \sum_{i \in M_p} \alpha_i = \sum_{i \in M_n} \alpha_i = 1, \alpha_i \geq 0 \right\}. \quad (9)
\end{aligned}$$

In Section 6, we present a rigorous proof that under some assumptions on $\ell(\xi)$, the min-max theorem works in the above Lagrangian function, i.e., there is no duality gap. For each binary label, we define the parametrized uncertainty sets, $\mathcal{U}_p[c]$ and $\mathcal{U}_n[c]$, by

$$o \in \{p, n\}, \quad \mathcal{U}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : \alpha_i \geq 0, \sum_{i \in M_o} \alpha_i = 1, \frac{1}{m} \sum_{i \in M_o} \ell^*(m\alpha_i) \leq c \right\}. \quad (10)$$

Then, the optimization problem in (9) is represented by

$$\inf_{c_p, c_n, \mathbf{z}_p, \mathbf{z}_n} c_p + c_n + \lambda \|\mathbf{z}_p - \mathbf{z}_n\| \quad \text{subject to} \quad \mathbf{z}_p \in \mathcal{U}_p[c_p], \mathbf{z}_n \in \mathcal{U}_n[c_n], c_p, c_n \in \mathbb{R}. \quad (11)$$

Let $\hat{\mathbf{z}}_p$ and $\hat{\mathbf{z}}_n$ be the optimal solution of \mathbf{z}_p and \mathbf{z}_n in (11). Let $\hat{\mathbf{w}}$ be an optimal solution of \mathbf{w} in (8). The saddle point of the above min-max problem (9) provides the relation between the $\hat{\mathbf{z}}_p$, $\hat{\mathbf{z}}_n$ and $\hat{\mathbf{w}}$. Some calculation yields that, when $\hat{\mathbf{z}}_p = \hat{\mathbf{z}}_n$ holds, any vector such that $\|\hat{\mathbf{w}}\|^2 \leq \lambda^2$ satisfies the KKT condition of (8). On the other hand, when $\hat{\mathbf{z}}_p \neq \hat{\mathbf{z}}_n$ holds, $\hat{\mathbf{w}}$ is given by $\hat{\mathbf{w}} = \lambda(\hat{\mathbf{z}}_p - \hat{\mathbf{z}}_n)/\|\hat{\mathbf{z}}_p - \hat{\mathbf{z}}_n\|$. Hence, an optimal solution of the normal vector in the linear decision function is given as

$$\hat{\mathbf{w}} = \begin{cases} \frac{\lambda}{\|\hat{\mathbf{z}}_p - \hat{\mathbf{z}}_n\|} (\hat{\mathbf{z}}_p - \hat{\mathbf{z}}_n), & \hat{\mathbf{z}}_p \neq \hat{\mathbf{z}}_n, \\ \mathbf{0}, & \hat{\mathbf{z}}_p = \hat{\mathbf{z}}_n. \end{cases} \quad (12)$$

We show a sufficient condition that the equality $\hat{\mathbf{z}}_p = \hat{\mathbf{z}}_n$ holds. Suppose that $\mathcal{U}_p[c_p] \cap \mathcal{U}_n[c_n]$ is nonempty for all c_p and c_n , whenever $\mathcal{U}_p[c_p]$ and $\mathcal{U}_n[c_n]$ are both nonempty. Then, clearly $\mathbf{z}_p = \mathbf{z}_n \in \mathcal{U}_p[c_p] \cap \mathcal{U}_n[c_n]$ is the optimal choice of the objective function in (11). In ν -SVM with a small $\nu > 0$, the reduced convex-hulls satisfy $\mathcal{U}_p \cap \mathcal{U}_n = \emptyset$, and hence, $\hat{\mathbf{z}}_p = \hat{\mathbf{z}}_n$ and $\hat{\mathbf{w}} = \mathbf{0}$ hold.

The bias term b in the linear decision function is not directly obtained from the optimal solution of (11) without knowing the explicit form of the loss function ℓ . A simple way of estimating the bias term is to choose $\hat{b} = -(\hat{\mathbf{w}}^T \hat{\mathbf{z}}_p + \hat{\mathbf{w}}^T \hat{\mathbf{z}}_n)/2$, which provides the decision

Learning with uncertainty set:

- Step 1.** Given training samples, we construct parametrized uncertainty sets $\mathcal{U}_p[c]$ and $\mathcal{U}_n[c]$ in some way.
- Step 2.** Solve (11), and obtain the normal vector by (12).
- Step 3.** The bias term of the decision function is estimated by (13).

Figure 2: Learning algorithm based on uncertainty set.

boundary bisecting the line segment connecting $\hat{\mathbf{z}}_p$ and $\hat{\mathbf{z}}_n$. In the learning algorithm proposed in Section 5, the bias term is estimated by minimizing the error rate

$$\min_{b \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^m \mathbb{I}[y_i(\hat{\mathbf{w}}^T \mathbf{x}_i + b) \leq 0]. \quad (13)$$

Since the estimated normal vector $\hat{\mathbf{w}}$ is substituted in the above objective function, the optimization is tractable.

Based on the argument above, we propose the learning algorithm using uncertainty sets in Figure 2. It is straightforward to apply the kernel method to the algorithm. In order to study statistical properties of the learning algorithm based on uncertainty sets, we need more elaborate description on the algorithm. Details are presented in Section 5.

We show some examples of uncertainty sets (10) associated with popular loss functions. In the following examples, the index sets, M_p and M_n , are defined by (6) for the training samples $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$, and let m_p and m_n be $m_p = |M_p|$ and $m_n = |M_n|$, respectively.

Example 1 (ν -SVM). *As explained above, the problem (8) is reduced to ν -SVM by defining $\ell(z) = \max\{2z/\nu, 0\}$. The conjugate function of ℓ is given as*

$$\ell^*(\alpha) = \begin{cases} 0, & \alpha \in [0, 2/\nu], \\ \infty, & \alpha \notin [0, 2/\nu], \end{cases}$$

and the associated uncertainty set is defined by

$$o \in \{p, n\}, \quad \mathcal{U}_o[c] = \begin{cases} \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : \sum_{i \in M_o} \alpha_i = 1, 0 \leq \alpha_i \leq \frac{2}{m\nu}, i \in M_o \right\}, & c \geq 0, \\ \emptyset, & c < 0. \end{cases}$$

For $c \geq 0$, the uncertainty set consists of the reduced convex-hull of training samples, and it does not depend on the parameter c . In addition, the negative c is infeasible. Hence, in the problem (11), optimal solutions of c_p and c_n are given as $c_p = c_n = 0$, and the problem is reduced to the simple minimum distance problem.

Example 2 (Truncated quadratic loss). Now consider $\ell(z) = (\max\{1 + z, 0\})^2$. The conjugate function is

$$\ell^*(\alpha) = \begin{cases} -\alpha + \frac{\alpha^2}{4}, & \alpha \geq 0, \\ \infty, & \alpha < 0. \end{cases}$$

For $o \in \{p, n\}$, we define $\bar{\mathbf{x}}_o$ and $\hat{\Sigma}_o$ as the empirical mean and the empirical covariance matrix of the samples $\{\mathbf{x}_i : i \in M_o\}$, i.e.,

$$\bar{\mathbf{x}}_o = \frac{1}{m_o} \sum_{i \in M_o} \mathbf{x}_i, \quad \hat{\Sigma}_o = \frac{1}{m_o} \sum_{i \in M_o} (\mathbf{x}_i - \bar{\mathbf{x}}_o)(\mathbf{x}_i - \bar{\mathbf{x}}_o)^T.$$

Suppose that $\hat{\Sigma}_o$ is invertible. Then, the uncertainty set corresponding to the truncated quadratic loss is given as

$$\begin{aligned} o \in \{p, n\}, \quad \mathcal{U}_o[c] &= \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0, i \in M_o, \sum_{i \in M_o} \alpha_i^2 \leq \frac{4(c+1)}{m} \right\} \\ &= \left\{ \mathbf{z} \in \text{conv}\{\mathbf{x}_i : i \in M_o\} : (\mathbf{z} - \bar{\mathbf{x}}_o)^T \hat{\Sigma}_o^{-1} (\mathbf{z} - \bar{\mathbf{x}}_o) \leq \frac{4(c+1)m_o}{m} \right\}. \end{aligned}$$

To prove the second equality, let us define the matrix $X = (\mathbf{x}_1, \dots, \mathbf{x}_{m_o}) \in \mathbb{R}^{d \times m_o}$. For $\boldsymbol{\alpha}_o = (\alpha_i)_{i \in M_o}$ satisfying the constraints, the equality $\mathbf{z} = \sum_{i \in M_o} \alpha_i \mathbf{x}_i = (X - \bar{\mathbf{x}}_o \mathbf{1}^T) \boldsymbol{\alpha}_o + \bar{\mathbf{x}}_o$ holds, where $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^{m_o}$. Then, the singular value decomposition of the matrix $X - \bar{\mathbf{x}}_o \mathbf{1}^T$ and the constraint $\|\boldsymbol{\alpha}_o\|^2 \leq 4(c+1)/m$ yield the second equality. A similar uncertainty set is used in minimax probability machine (MPM) [15] and maximum margin MPM [18], though the constraint, $\mathbf{z} \in \text{conv}\{\mathbf{x}_i : i \in M_o\}$, is not imposed in these learning methods.

Example 3 (exponential loss). The loss function $\ell(z) = e^z$ is used in Adaboost [12, 13]. The conjugate function is equal to

$$\ell^*(\alpha) = \begin{cases} -\alpha + \alpha \log \alpha, & \alpha \geq 0, \\ \infty, & \alpha < 0. \end{cases}$$

Hence, the corresponding uncertainty set is defined as

$$\mathcal{U}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0, i \in M_o, \sum_{i \in M_o} \alpha_i \log \frac{\alpha_i}{1/m_o} \leq c + 1 + \log \frac{m_o}{m} \right\}$$

for $o \in \{p, n\}$. In the uncertainty set, the Kullback-Leibler divergence from the weight $\alpha_i, i \in M_o$ to the uniform weight is bounded above.

In this section, we derived parametrized uncertainty sets associated with convex loss functions. Inversely, if the uncertainty set is represented as the form of (10), there exists the corresponding loss function. When we consider statistical properties of the classifier estimated based on the uncertainty set, we can study the equivalent estimator derived from the corresponding loss function. We have many theoretical tools to analyze such estimators. However, if the uncertainty set does not have the expression of (10), the corresponding loss function would not exist. In this case, we cannot apply the standard theoretical tools to understand statistical properties of learning algorithms based on such uncertainty sets. One way to remedy the drawback is to revise the uncertainty set so as to possess the corresponding loss function. The next section is devoted to study a way of revising the uncertainty set.

4 Revision of Uncertainty Sets

Given a parametrized uncertainty set, generally there does not exist the loss function which corresponds to the uncertainty set. In this section, we present a way of revising the uncertainty set such that there exists a corresponding loss function.

We consider two kinds of representations for parametrized uncertainty sets: one is vertex representation, and the other is level-set representation. Let M_p and M_n be index sets defined in (6), and we define $m_p = |M_p|$ and $m_n = |M_n|$. For $o \in \{p, n\}$, let L_o be a closed, convex, proper function on \mathbb{R}^{m_o} , and L_o^* be the conjugate function of L_o . The argument of L_o^* is represented by $\alpha_o = (\alpha_i)_{i \in M_o}$. The *vertex representation* of the uncertainty set is defined as

$$\mathcal{U}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : L_o^*(\alpha_o) \leq c \right\}, \quad o \in \{p, n\}. \quad (14)$$

In Example 2, the function $L_o^*(\alpha_o) = \frac{m}{4} \sum_{i \in M_o} \alpha_i^2 - 1$ is employed. On the other hand, let us define $h_o : \mathbb{R}^d \rightarrow \mathbb{R}$ as a closed, convex, proper function, and h_o^* be the conjugate of h_o . The *level-set representation* of the uncertainty set is defined by

$$\mathcal{U}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : h_o^* \left(\sum_{i \in M_o} \alpha_i \mathbf{x}_i \right) \leq c \right\}, \quad o \in \{p, n\}. \quad (15)$$

The function h_o^* may depend on the population distribution. We suppose that h_o^* does not depend on the sample points, $\mathbf{x}_i, i \in M_o$. In Example 2, the second expression of the uncertainty set involves the convex function $h_o^*(\mathbf{z}) = (\mathbf{z} - \bar{\mathbf{x}}_o)^T \hat{\Sigma}_o^{-1} (\mathbf{z} - \bar{\mathbf{x}}_o)$. This function does not satisfy the assumption, since h_o^* depends on training samples via $\bar{\mathbf{x}}_o$ and $\hat{\Sigma}_o$. Instead, the function $h_o^*(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\mu}_o)^T \Sigma_o^{-1} (\mathbf{z} - \boldsymbol{\mu}_o)$ with the population mean $\boldsymbol{\mu}_o$ and the population covariance matrix Σ_o meets the condition. When $\boldsymbol{\mu}_o$ and Σ_o are replaced with the estimated parameters based on a prior knowledge or a set of samples independent of the training samples, $\{\mathbf{x}_i : i \in M_o\}$, the function h_o^* with the estimated parameters still satisfies the condition we imposed above.

4.1 From uncertainty sets to loss functions

In popular learning algorithms using uncertainty sets such as hard-margin SVM, ν -SVM and maximum margin MPM, the decision function is estimated by solving the minimum distance problem (4) with $\mathcal{U}_p = \mathcal{U}_p[\bar{c}_p]$ and $\mathcal{U}_n = \mathcal{U}_n[\bar{c}_n]$, where \bar{c}_p and \bar{c}_n are prespecified constants. In order to investigate the statistical properties of the learning algorithm using uncertainty sets, we consider the primal expression of a variant of the minimum distance problem (4).

In Section 3, we derived the problem (11) as the dual form of (8). Here, we consider the following optimization problem to obtain the loss function corresponding to given uncertainty sets having the vertex representation (14),

$$\begin{aligned} \min_{c_p, c_n, \mathbf{z}_p, \mathbf{z}_n} \quad & c_p + c_n + \lambda \|\mathbf{z}_p - \mathbf{z}_n\| \\ \text{subject to} \quad & c_p, c_n \in \mathbb{R}, \\ & \mathbf{z}_p \in \mathcal{U}_p[c_p] \cap \text{conv}\{\mathbf{x}_i : i \in M_p\}, \\ & \mathbf{z}_n \in \mathcal{U}_n[c_n] \cap \text{conv}\{\mathbf{x}_i : i \in M_n\}. \end{aligned} \quad (16)$$

In the above problem the constraints, $\mathbf{z}_o \in \text{conv}\{\mathbf{x}_i : i \in M_o\}, o \in \{p, n\}$, are added, since the corresponding uncertainty set (10) has the same constraint. We derive the primal problem

corresponding to (16) via the min-max theorem. A brief calculation yields that (16) is equivalent to

$$\begin{aligned} \min_{\alpha} \quad & L_p^*(\alpha_p) + L_n^*(\alpha_n) + \lambda \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\| \\ \text{subject to} \quad & \sum_{i \in M_p} \alpha_i = 1, \sum_{j \in M_n} \alpha_j = 1, \alpha_i \geq 0 \ (i = 1, \dots, m). \end{aligned} \quad (17)$$

If there is no duality gap, the corresponding primal formulation of (17) is given as

$$\begin{aligned} \inf_{\mathbf{w}, b, \rho, \xi_p, \xi_n} \quad & -2\rho + L_p(\xi_p) + L_n(\xi_n), \\ \text{subject to} \quad & \rho - y_i(\mathbf{w}^T \mathbf{x}_i + b) \leq \xi_i, \ i = 1, \dots, m, \quad \|\mathbf{w}\|^2 \leq \lambda^2, \end{aligned} \quad (18)$$

where ξ_o is defined as $\xi_o = (\xi_i)_{i \in M_o}$ for $o \in \{p, n\}$.

In the primal expression (18), L_p and L_n are regarded as the loss function for the decision function $\mathbf{w}^T \mathbf{x} + b$ on training samples. In general, however, the loss function is not represented as the empirical mean over training samples. Thus, we cannot apply the standard theoretical tools to investigate statistical properties such as Bayes risk consistency for the learning algorithm based on (16) or (18). On the other hand, if the problem (18) is described as the empirical loss minimization, we can study statistical properties of the algorithm by applying the statistical theory developed by [29, 26, 2]. To link the uncertainty set approach with the empirical loss minimization, we consider a revision of the uncertainty set.

4.2 Revised uncertainty sets and corresponding loss functions

We propose a way of revising uncertainty sets such that the primal form (18) is represented as minimization of the empirical mean of a loss function. Remember that the additivity of the function is kept unchanged in the conjugate function, i.e., $(\ell_1(z_1) + \ell_2(z_2))^* = (\ell_1(z_1))^* + (\ell_2(z_2))^*$.

Revision of uncertainty set defined by vertex representation: Suppose that the uncertainty set is described by (14). For $o \in \{p, n\}$, we define m_o -dimensional vectors $\mathbf{1}_o = (1, \dots, 1)$ and $\mathbf{0}_o = (0, \dots, 0)$. For the convex function $L_o^* : \mathbb{R}^{m_o} \rightarrow \mathbb{R}$, we define $\bar{\ell}^* : \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$ by

$$\bar{\ell}^*(\alpha) = \begin{cases} L_p^*\left(\frac{\alpha}{m} \mathbf{1}_p\right) + L_n^*\left(\frac{\alpha}{m} \mathbf{1}_n\right) - L_p^*(\mathbf{0}_p) - L_n^*(\mathbf{0}_n) & \alpha \geq 0, \\ \infty, & \alpha < 0. \end{cases} \quad (19)$$

The revised uncertainty set $\bar{\mathcal{U}}_o[c]$, $o \in \{p, n\}$ is defined as

$$\bar{\mathcal{U}}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0, i \in M_o, \frac{1}{m} \sum_{i \in M_o} \bar{\ell}^*(\alpha_i m) \leq c \right\}.$$

Revision of uncertainty set defined by level-set representation: Suppose that the uncertainty set is described by (15) and that the mean of the input vector \mathbf{x} conditioned on the positive (resp. negative) label is given as $\boldsymbol{\mu}_p$ (resp. $\boldsymbol{\mu}_n$). The null vector is denoted as $\mathbf{0}$. We define the function $\bar{\ell}^* : \mathbb{R} \rightarrow \mathbb{R}$ by

$$\bar{\ell}^*(\alpha) = \begin{cases} h_p^*\left(\alpha \frac{m_p}{m} \boldsymbol{\mu}_p\right) + h_n^*\left(\alpha \frac{m_n}{m} \boldsymbol{\mu}_n\right) - h_p^*(\mathbf{0}) - h_n^*(\mathbf{0}) & \alpha \geq 0, \\ \infty, & \alpha < 0. \end{cases} \quad (20)$$

The revised uncertainty set $\bar{\mathcal{U}}_o[c]$, $o \in \{p, n\}$ is defined as

$$\bar{\mathcal{U}}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0, i \in M_o, \frac{1}{m} \sum_{i \in M_o} \bar{\ell}^*(\alpha_i m) \leq c, \right\}.$$

We apply the parallel shift of training samples so as to be $\boldsymbol{\mu}_p \neq \mathbf{0}$ or $\boldsymbol{\mu}_n \neq \mathbf{0}$.

We explain the reason why the revised uncertainty set is defined as above. In the revision (19), the uncertainty set is kept unchanged, when the function $L_p^* + L_n^*$ is described in the additive form. The precise description is presented in the following theorem.

Theorem 1. *Let $L_o^* : \mathbb{R}^{m_o} \rightarrow \mathbb{R}$, $o \in \{p, n\}$ be convex functions, and $\bar{\ell}^*$ be the function defined by (19) for given L_p^* and L_n^* . Suppose that $\ell : \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$ is a closed, convex, proper function such that $\ell^*(0) = 0$ and $\ell^*(\alpha) = \infty$ for $\alpha < 0$ hold.*

1. *Suppose that the equality*

$$L_p^*(\boldsymbol{\alpha}_p) + L_n^*(\boldsymbol{\alpha}_n) - L_p^*(\mathbf{0}_p) - L_n^*(\mathbf{0}_n) = \frac{1}{m} \sum_{i=1}^m \ell^*(\alpha_i m)$$

holds for all non-negative α_i , $i = 1, \dots, m$. Then, the equality $\bar{\ell}^ = \ell^*$ holds.*

2. *Suppose that the equality*

$$L_p^*(\alpha \mathbf{1}_p) + L_n^*(\alpha \mathbf{1}_n) - L_p^*(\mathbf{0}_p) - L_n^*(\mathbf{0}_n) = \frac{1}{m} \sum_{i=1}^m \ell^*(\alpha m) = \ell^*(\alpha m)$$

holds for all $\alpha \geq 0$. Then, the equality $\bar{\ell}^ = \ell^*$ holds.*

Proof. We prove the first statement. From the definition of $\bar{\ell}^*$ and the assumption on ℓ^* , the equality $\ell^*(\alpha) = \bar{\ell}^*(\alpha)$ holds for $\alpha < 0$. Suppose $\alpha \geq 0$. The assumption on L_p^* and L_n^* leads to $L_p^*(\frac{\alpha}{m} \mathbf{1}_p) + L_n^*(\frac{\alpha}{m} \mathbf{1}_n) - L_p^*(\mathbf{0}_p) - L_n^*(\mathbf{0}_n) = \ell^*(\alpha)$. Hence, we have $\ell^* = \bar{\ell}^*$. The second statement of the theorem is straightforward. \square

Theorem 1 implies that the transformation of $L_p^* + L_n^*$ to $\frac{1}{m} \sum_{i=1}^m \bar{\ell}^*(\alpha_i m)$ is a projection onto the set of functions with the additive form. In addition, the second statement of Theorem 1 denotes that the projection is uniquely determined when we impose the condition that the values on the diagonal $\{(\alpha, \dots, \alpha) \in \mathbb{R}^m : \alpha \geq 0\}$ are unchanged.

Next, we explain the validity of the formula (20). We want to find a function $\bar{\ell}^*(\alpha)$ such that $h_p^*(\sum_{i \in M_p} \alpha_i \mathbf{x}_i) + h_n^*(\sum_{i \in M_n} \alpha_i \mathbf{x}_i) - h_p^*(\mathbf{0}) - h_n^*(\mathbf{0})$ is close to $\frac{1}{m} \sum_{i=1}^m \bar{\ell}^*(m \alpha_i)$ in some sense. We substitute $\alpha_i = \alpha/m$ into $h_o^*(\sum_{i \in M_o} \alpha_i \mathbf{x}_i)$, $o \in \{p, n\}$. In the large sample limit, $h_o^*(\sum_{i \in M_o} \alpha/m \mathbf{x}_i)$ is approximated by $h_o^*(\alpha \frac{m_o}{m} \boldsymbol{\mu}_o)$. Suppose that

$$h_p^*(\alpha \frac{m_p}{m} \boldsymbol{\mu}_p) + h_n^*(\alpha \frac{m_n}{m} \boldsymbol{\mu}_n) - h_p^*(\mathbf{0}) - h_n^*(\mathbf{0})$$

is represented as $\frac{1}{m} \sum_{i=1}^m \bar{\ell}^*(\frac{\alpha}{m} m) = \bar{\ell}^*(\alpha)$. Then, we obtain (20).

For the revised uncertainty sets $\bar{\mathcal{U}}_p[c]$ and $\bar{\mathcal{U}}_n[c]$, the corresponding primal problem of

$$\min_{c_p, c_n, \mathbf{z}_p, \mathbf{z}_n} c_p + c_n + \lambda \|\mathbf{z}_p - \mathbf{z}_n\| \quad \text{subject to } \mathbf{z}_p \in \bar{\mathcal{U}}_p[c_p], \mathbf{z}_n \in \bar{\mathcal{U}}_n[c_n] \quad (21)$$

is given as

$$\inf_{w, b, \rho, \xi_p, \xi_n} -2\rho + \frac{1}{m} \sum_{i=1}^m \bar{\ell}(\xi_i) \quad \text{subject to } \rho - y_i(\mathbf{w}^T \mathbf{x}_i + b) \leq \xi_i, i = 1, \dots, m, \quad \|\mathbf{w}\|^2 \leq \lambda^2.$$

The revision of the uncertainty sets leads to the empirical mean of the revised loss function $\bar{\ell}$. When we study statistical properties of the estimator given by the optimal solution of (21), we can apply the standard theoretical tools, since the objective in the primal expression is described by the empirical mean of the revised loss functions.

We show some examples to illustrate how the revision of the uncertainty set works.

Example 4. Let L_o^* , $o \in \{p, n\}$ be the convex function $L_o^*(\boldsymbol{\alpha}_o) = \boldsymbol{\alpha}_o^T C_o \boldsymbol{\alpha}_o$, where C_o is a positive definite matrix. The revised function defined by (19) is given as

$$\bar{\ell}^*(\alpha) = \alpha^2 \frac{\mathbf{1}_p^T C_p \mathbf{1}_p + \mathbf{1}_n^T C_n \mathbf{1}_n}{m^2}$$

for $\alpha \geq 0$. Then, we have

$$\frac{1}{m} \sum_{i=1}^m \bar{\ell}^*(\alpha_i m) = \frac{\mathbf{1}_p^T C_p \mathbf{1}_p + \mathbf{1}_n^T C_n \mathbf{1}_n}{m} \sum_{i=1}^m \alpha_i^2$$

When both C_p and C_n are the identity matrix, the equality

$$L_p^*(\boldsymbol{\alpha}_p) + L_n^*(\boldsymbol{\alpha}_n) = \frac{1}{m} \sum_{i=1}^m \bar{\ell}^*(\alpha_i m) = \sum_{i=1}^m \alpha_i^2$$

holds. Let k be $k = \mathbf{1}_p^T C_p \mathbf{1}_p + \mathbf{1}_n^T C_n \mathbf{1}_n$. Then, the revised uncertainty set is given as

$$o \in \{p, n\}, \quad \bar{\mathcal{U}}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0 (i \in M_o), \sum_{i \in M_o} \alpha_i^2 \leq \frac{cm}{k} \right\}.$$

For $o \in \{p, n\}$, let $\bar{\mathbf{x}}_o$ and $\hat{\Sigma}_o$ be the empirical mean and the empirical covariance matrix,

$$\bar{\mathbf{x}}_o = \frac{1}{m_o} \sum_{i \in M_o} \mathbf{x}_i, \quad \hat{\Sigma}_o = \frac{1}{m_o} \sum_{i \in M_o} (\mathbf{x}_i - \bar{\mathbf{x}}_o)(\mathbf{x}_i - \bar{\mathbf{x}}_o)^T.$$

If $\hat{\Sigma}_o$ is invertible, we have

$$\bar{\mathcal{U}}_o[c] = \left\{ \mathbf{z} \in \text{conv}\{\mathbf{x}_i : i \in M_o\} : (\mathbf{z} - \bar{\mathbf{x}}_o)^T \hat{\Sigma}_o^{-1} (\mathbf{z} - \bar{\mathbf{x}}_o) \leq \frac{cm m_o}{k} \right\}.$$

In the learning algorithm based on the revised uncertainty set, the estimator is obtained by solving

$$\begin{aligned} & \min_{c_p, c_n, \mathbf{z}_p, \mathbf{z}_n} c_p + c_n + \lambda \|\mathbf{z}_p - \mathbf{z}_n\| \quad \text{subject to } \mathbf{z}_p \in \bar{\mathcal{U}}_p[c_p], \mathbf{z}_n \in \bar{\mathcal{U}}_n[c_n] \\ \iff & \min_{c_p, c_n, \mathbf{z}_p, \mathbf{z}_n} c_p + c_n + \frac{m^2 \lambda}{4k} \|\mathbf{z}_p - \mathbf{z}_n\| \quad \text{subject to } \mathbf{z}_p \in \bar{\mathcal{U}}_p \left[\frac{4c_p k}{m^2} \right], \mathbf{z}_n \in \bar{\mathcal{U}}_n \left[\frac{4c_n k}{m^2} \right]. \end{aligned}$$

The corresponding primal expression is given as

$$\min_{w, b, \rho, \xi} -2\rho + \frac{1}{m} \sum_{i \in M_p} \xi_i^2 \quad \text{subject to } \rho - y_i(\mathbf{w}^T \mathbf{x}_i + b) \leq \xi_i, 0 \leq \xi_i, \forall i, \quad \|\mathbf{w}\|^2 \leq \left(\frac{m^2 \lambda}{4k} \right)^2.$$

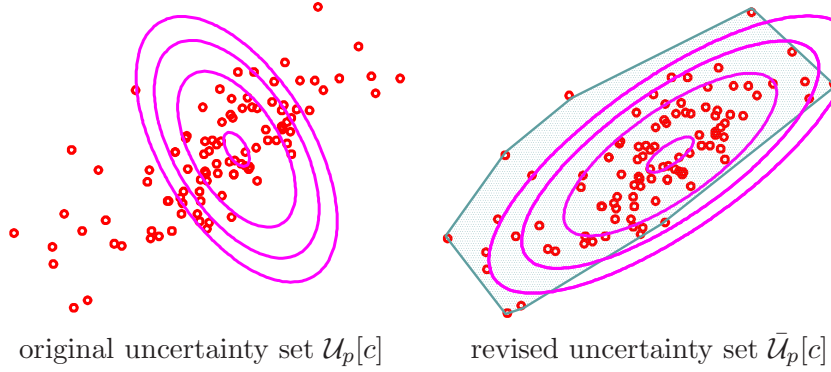


Figure 3: Training samples and the uncertainty sets are depicted. Left panel: the original uncertainty set for the positive label. Right panel: the revised uncertainty set which consists of the intersection of the ellipsoid and the convex-hull of the input vectors with positive label.

Example 5. We define $h_o^* : \mathcal{X} \rightarrow \mathbb{R}$ for $o \in \{p, n\}$ by

$$h_o^*(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\mu}_o)^T C_o (\mathbf{z} - \boldsymbol{\mu}_o)$$

where $\boldsymbol{\mu}_o$ is the mean vector of the input vector \mathbf{x} conditioned on each label and C_o is a positive definite matrix. In practice, the mean vector is estimated by using a prior knowledge which is independent of the training samples $\{(\mathbf{x}_i, y_i) : i = 1, \dots, m\}$. Suppose that $\boldsymbol{\mu}_o \neq \mathbf{0}$. Then, for $\alpha \geq 0$, the revision of (20) leads to

$$\begin{aligned} \bar{\ell}^*(\alpha) &= \left(\left(\alpha \frac{m_p}{m} - 1 \right)^2 - 1 \right) \boldsymbol{\mu}_p^T C_p \boldsymbol{\mu}_p + \left(\left(\alpha \frac{m_n}{m} - 1 \right)^2 - 1 \right) \boldsymbol{\mu}_n^T C_n \boldsymbol{\mu}_n \\ &= b_1 \alpha + b_2 \alpha^2, \end{aligned}$$

where b_1 and $b_2 (> 0)$ are constant numbers. Thus, we have

$$\begin{aligned} \bar{\mathcal{U}}_o[c] &= \left\{ \sum_{i \in M_o} \alpha_i \mathbf{x}_i : \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0 (i \in M_o), \sum_{i \in M_o} \alpha_i^2 \leq \frac{c - b_1}{mb_2} \right\} \\ &= \left\{ \mathbf{z} \in \text{conv}\{\mathbf{x}_i : i \in M_o\} : (\mathbf{z} - \bar{\mathbf{x}}_o)^T \hat{\Sigma}_o^{-1} (\mathbf{z} - \bar{\mathbf{x}}_o) \leq m_o \cdot \frac{c - b_1}{mb_2} \right\}, \end{aligned}$$

where $\bar{\mathbf{x}}_o$ and $\hat{\Sigma}_o$ are the estimators of the mean vector and the covariance matrix based on training samples $\{\mathbf{x}_i : i \in M_o\}$. The corresponding loss function is obtained in the same way as Example 4. Figure 3 illustrates an example of the revision of the uncertainty set. In the left panel, the uncertainty set does not match the distribution of the training samples. The revised uncertainty set in the right panel seems to well approximate the dispersal of the training samples.

Example 6. We suppose that for $o \in \{p, n\}$, $\boldsymbol{\mu}_o$ is the mean vector and Σ_o is the covariance matrix of the input vector conditioned on each label. We define the uncertainty set by

$$o \in \{p, n\}, \quad \mathcal{U}_o[c] = \left\{ \mathbf{z} \in \text{conv}\{\mathbf{x}_i : i \in M_o\} : (\mathbf{z} - \boldsymbol{\mu})^T \Sigma_o^{-1} (\mathbf{z} - \boldsymbol{\mu}) \leq c, \forall \boldsymbol{\mu} \in \mathcal{A} \right\},$$

where \mathcal{A} denotes the estimation error of the mean vector $\boldsymbol{\mu}$. For a fixed radius $r > 0$, \mathcal{A} is defined as

$$\mathcal{A} = \{\boldsymbol{\mu} \in \mathcal{X} : (\boldsymbol{\mu} - \boldsymbol{\mu}_o)^T \Sigma_o^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_o) \leq r^2\}.$$

The uncertainty set with estimation error is used by [15] in MPM. The above uncertainty sets will be useful, when the probability in the training phase is slightly different from that in the test phase. Brief calculation yields that $\mathcal{U}_o[c]$ is represented by the level set of the convex function

$$h_o^*(\mathbf{z}) = \max_{\boldsymbol{\mu} \in \mathcal{A}} (\mathbf{z} - \boldsymbol{\mu})^T \Sigma_o^{-1} (\mathbf{z} - \boldsymbol{\mu}) = \left(\sqrt{(\mathbf{z} - \boldsymbol{\mu}_o)^T \Sigma_o^{-1} (\mathbf{z} - \boldsymbol{\mu}_o)} + r \right)^2$$

The revised uncertainty set $\bar{\mathcal{U}}_o[c]$ is defined by the function $\bar{\ell}^*$ which is given as

$$\begin{aligned} \bar{\ell}^*(\alpha) = & \left(\left| \alpha \frac{m_p}{m} - 1 \right| \sqrt{\boldsymbol{\mu}_p^T \Sigma_p^{-1} \boldsymbol{\mu}_p} + r \right)^2 - \left(\sqrt{\boldsymbol{\mu}_p^T \Sigma_p^{-1} \boldsymbol{\mu}_p} + r \right)^2 \\ & + \left(\left| \alpha \frac{m_n}{m} - 1 \right| \sqrt{\boldsymbol{\mu}_n^T \Sigma_n^{-1} \boldsymbol{\mu}_n} + r \right)^2 - \left(\sqrt{\boldsymbol{\mu}_n^T \Sigma_n^{-1} \boldsymbol{\mu}_n} + r \right)^2. \end{aligned} \quad (22)$$

We suppose that $\boldsymbol{\mu}_p \neq \mathbf{0}$ and $\boldsymbol{\mu}_n = \mathbf{0}$ hold. Let $d = \sqrt{\boldsymbol{\mu}_p^T \Sigma_p^{-1} \boldsymbol{\mu}_p}$ and $h = r/d (> 0)$. Then, the corresponding loss function is given as

$$\bar{\ell}(z) = \frac{md^2}{m_p} u\left(\frac{z}{d^2}\right),$$

where $u(z)$ as defined as

$$u(z) = \begin{cases} 0, & z \leq -2h - 2, \\ \left(\frac{z}{2} + 1 + h\right)^2, & -2h - 2 \leq z \leq -2h, \\ z + 2h + 1, & -2h \leq z \leq 2h, \\ \frac{z^2}{4} + z(1 - h) + (1 + h)^2, & 2h \leq z. \end{cases} \quad (23)$$

Figure 4 depicts the function $u(z)$ with $h = 1$. When $r = 0$ holds, $\bar{\ell}(z)$ is reduced to the truncated quadratic function shown in Example 4 and 5. For positive r , $\bar{\ell}(z)$ is linear around $z = 0$. This implies that by introducing the confidence set of the mean vector, \mathcal{A} , the penalty for the misclassification is reduced from quadratic to linear around the decision boundary, though the original uncertainty set $\mathcal{U}_o[c]$ does not correspond to minimization of an empirical loss function.

5 Kernel-based Learning Algorithm

We present a kernel variant of the learning algorithm using uncertainty sets. Suppose that training samples $(x_1, y_1), \dots, (x_m, y_m) \in \mathcal{X} \times \{+1, -1\}$ are observed, where \mathcal{X} is not necessarily a linear space. We define the kernel function $k : \mathcal{X}^2 \rightarrow \mathbb{R}$, and let \mathcal{H} be the reproducing kernel Hilbert space (RKHS) endowed with the kernel function k . See [23] for the details of the kernel estimators in machine learning. We consider the estimator of the decision function having the

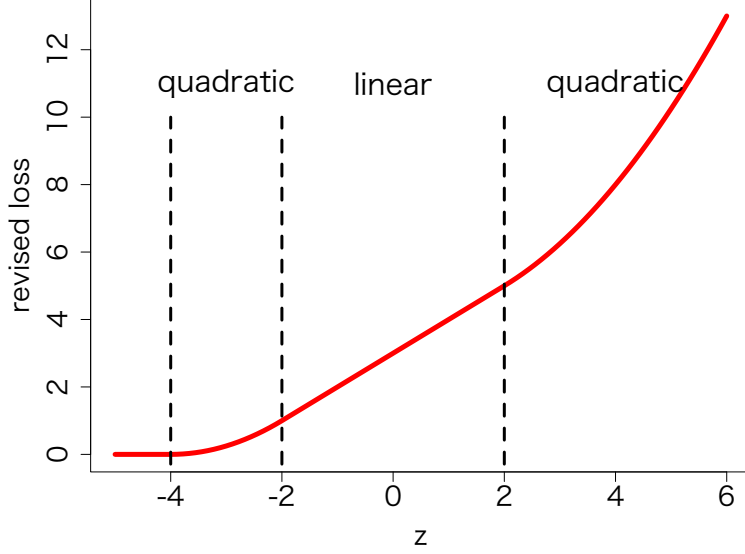


Figure 4: The loss function $u(z)$ in Example 6 is depicted, which corresponds to the revised uncertainty set with the estimation error.

form of $f(x) + b$, where $f \in \mathcal{H}$, $b \in \mathbb{R}$. In our algorithm, the function part $f(x)$ and the bias term b are separately estimated.

Figure 5 shows a kernel variant of the learning algorithm based on uncertainty sets. The algorithm is regarded as an extension of ν -SVM and maximum margin MPM, since the uncertainty set is extended from reduced convex-hull or ellipsoidal uncertainty set to general uncertainty set. The proposed algorithm is also a revision of the existing method based on the simple minimum distance problem. We shall illustrate the proposed algorithm in the below.

In the learning algorithm, training samples are divided into two disjoint subsets, T_1 and T_2 , which are described as

$$T_k = \{(x_i^{(k)}, y_i^{(k)}) : i = 1, \dots, m_k\}, \quad k = 1, 2.$$

The reason that we decompose the training samples is to simplify the analysis of statistical properties of the learning algorithm. In the kernel-based algorithm, the uncertainty sets, $\mathcal{U}_p[c]$ and $\mathcal{U}_n[c]$, are convex subsets in \mathcal{H} . Let M_p and M_n be the index sets of T_1 defined by

$$M_p = \{i : y_i^{(1)} = +1, i = 1, \dots, m_1\}, \quad M_n = \{i : y_i^{(1)} = -1, i = 1, \dots, m_1\}.$$

For $o \in \{p, n\}$, the uncertainty set $\mathcal{U}_o[c] \subset \mathcal{H}$ is defined as a convex subset of the convex-hull of $\{k(\cdot, x_i^{(1)}) : i \in M_o\}$. Moreover, we assume that the monotonicity $\mathcal{U}_o[c] \subset \mathcal{U}_o[c']$ holds for $c \leq c'$. If necessary, we revise the uncertainty set as shown in Section 4 in order to link the uncertainty set with a loss function.

When the uncertainty sets involve some parameters to be estimated, a prior knowledge or additional samples independent of the training samples $T_1 \cup T_2$ are used for its estimation. For example, the uncertainty set defined by the level set of $h_o(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\mu}_o)^T \Sigma_o^{-1} (\mathbf{z} - \boldsymbol{\mu}_o)$, $o \in \{p, n\}$

Inputs. Decompose the training samples into two disjoint subsets,

$$T_1 = \{(x_i^{(1)}, y_i^{(1)}) : i = 1, \dots, m_1\}, \quad T_2 = \{(x_i^{(2)}, y_i^{(2)}) : i = 1, \dots, m_2\}.$$

For the set of training samples T_1 , let M_p and M_n be the index sets defined by $M_p = \{i : y_i^{(1)} = +1, i = 1, \dots, m_1\}$ and $M_n = \{i : y_i^{(1)} = -1, i = 1, \dots, m_1\}$, respectively.

Initialization. We define the RKHS \mathcal{H} with the kernel function $k(x, x')$. Prepare the parametrized uncertainty sets $\mathcal{U}_p[c]$ and $\mathcal{U}_n[c]$ in \mathcal{H} such that

$$\mathcal{U}_p[c] \subset \text{conv}\{k(\cdot, x_i^{(1)}) : i \in M_p\}, \quad \mathcal{U}_n[c] \subset \text{conv}\{k(\cdot, x_i^{(1)}) : i \in M_n\}.$$

When the uncertainty sets involve some parameters to be estimated, a prior knowledge or additional samples independent of the training samples $T_1 \cup T_2$ are used for its estimation. If necessary, we apply the revision of the uncertainty sets presented in Section 4 in order to link the uncertainty set with a loss function. Set the regularization parameter $\lambda > 0$.

Step 1. Solve the optimization problem,

$$\begin{aligned} & \inf_{c_p, c_n, f_p, f_n} c_p + c_n + \lambda \|f_p - f_n\|_{\mathcal{H}} \\ & \text{subject to } f_p \in \mathcal{U}_p[c_p], \quad f_n \in \mathcal{U}_n[c_n], \quad c_p, c_n \in \mathbb{R}. \end{aligned}$$

Optimal solutions of f_p and f_n are denoted as \hat{f}_p and \hat{f}_n . Define \hat{f} by

$$\hat{f} = \begin{cases} \frac{\lambda}{\|\hat{f}_p - \hat{f}_n\|_{\mathcal{H}}} (\hat{f}_p - \hat{f}_n), & \hat{f}_p \neq \hat{f}_n, \\ 0, & \hat{f}_p = \hat{f}_n. \end{cases}$$

Step 2. Solve the one-dimensional optimization problem defined from the estimator \hat{f} and the data set T_2 ,

$$\min_{b \in \mathbb{R}} \hat{\mathcal{E}}_{T_2}(\hat{f} + b)$$

The optimal solution is denoted as \tilde{b} .

Output. The estimator of the decision function is given by $\hat{f}(x) + \tilde{b}$.

Figure 5: Kernel-based learning algorithm using uncertainty sets.

involves the mean vector $\boldsymbol{\mu}_o$ and the covariance matrix Σ_o . In our algorithm, we need to prepare additional samples to estimate $\boldsymbol{\mu}_o$ and Σ_o .

The subset T_1 is used for the estimation of the function part $f \in \mathcal{H}$ in the decision function. First, we solve the problem,

$$\begin{aligned} \inf_{c_p, c_n, f_p, f_n} \quad & c_p + c_n + \lambda \|f_p - f_n\|_{\mathcal{H}} \\ \text{subject to} \quad & f_p \in \mathcal{U}_p[c_p], \quad f_n \in \mathcal{U}_n[c_n], \quad c_p, c_n \in \mathbb{R}. \end{aligned} \quad (24)$$

Let \hat{f}_p and \hat{f}_n be optimal solutions of f_p and f_n in (24). Then, in the same way as (12), the function part of the decision function is estimated by

$$\hat{f} = \begin{cases} \frac{\lambda}{\|\hat{f}_p - \hat{f}_n\|_{\mathcal{H}}} (\hat{f}_p - \hat{f}_n), & \hat{f}_p \neq \hat{f}_n, \\ 0, & \hat{f}_p = \hat{f}_n. \end{cases} \quad (25)$$

For the estimation of the bias term b , the data set T_2 is used. The bias estimator \tilde{b} is an optimal solution of

$$\min_{b \in \mathbb{R}} \hat{\mathcal{E}}_{T_2}(\hat{f} + b). \quad (26)$$

Our purpose is to obtain the decision function with a low prediction error. Hence, the error rate (26) is an appropriate criterion for the estimation of the bias term. Though generally the minimization of the training error rate is hard task, the one-dimensional optimization is easily conducted. Then, the estimator of the decision function is given by $\hat{f}(x) + \tilde{b}$. By separating the training data used in Step 1 and Step 2, we can simplify the statistical analysis of the estimator.

6 Statistical Properties of Kernel-based Learning Algorithm

In this section, we study statistical properties of the learning algorithm presented in Figure 5. Especially, we prove that the expected 0-1 loss of the estimator, $\mathcal{E}(\hat{f} + \tilde{b})$, converges to the Bayes risk \mathcal{E}^* defined by (2).

6.1 Definitions and assumptions

We derive the dual representation of the learning algorithm in Figure 5. For a convex function $\ell : \mathbb{R} \rightarrow \mathbb{R}$, let ℓ^* be the conjugate function of ℓ . For $o \in \{p, n\}$, suppose that the uncertainty sets are described as the form of

$$\mathcal{U}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i k(\cdot, x_i^{(1)}) \in \mathcal{H} : \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0 (i \in M_o), \frac{1}{m} \sum_{i \in M_o} \ell^*(m\alpha_i) \leq c \right\}. \quad (27)$$

In the same way as the derivation in Section 3.2, we find that the problem (24) is the dual representation of

$$\begin{aligned} \min_{f, b, \rho} \quad & -2\rho + \frac{1}{m_1} \sum_{i=1}^{m_1} \ell(\rho - y_i^{(1)}(f(x_i^{(1)}) + b)) \\ \text{subject to} \quad & f \in \mathcal{H}, \quad b \in \mathbb{R}, \quad \rho \in \mathbb{R}, \quad \|f\|_{\mathcal{H}}^2 \leq \lambda^2. \end{aligned} \quad (28)$$

Later on, we show a rigorous proof of the duality between (28) and (24) with the uncertainty set (27). In order to investigate statistical properties of the learning algorithm using uncertainty sets, we consider the primal problem (28) and (26) instead of the dual problem (24) and (26).

We define some notations. For a measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$ and a real number $\rho \in \mathbb{R}$, we define the expected loss $\mathcal{R}(f, \rho)$ and the regularized expected loss $\mathcal{R}_\lambda(f, \rho)$ by

$$\begin{aligned}\mathcal{R}(f, \rho) &= -2\rho + \mathbb{E}[\ell(\rho - yf(x))], \\ \mathcal{R}_\lambda(f, \rho) &= -2\rho + \mathbb{E}[\ell(\rho - yf(x))] + \theta(\|f\|_{\mathcal{H}}^2 \leq \lambda^2),\end{aligned}$$

where λ is a positive number and $\theta(A)$ equals 0 when A is true and ∞ otherwise. Let \mathcal{R}^* be the infimum of $\mathcal{R}(f, \rho)$,

$$\mathcal{R}^* = \inf\{\mathcal{R}(f, \rho) : f \in L_0, \rho \in \mathbb{R}\}.$$

For the set of training samples, $T = \{(x_1, y_1), \dots, (x_m, y_m)\}$, the empirical loss $\widehat{\mathcal{R}}_T(f, \rho)$ and the regularized empirical loss $\widehat{\mathcal{R}}_{T, \lambda}(f, \rho)$ are defined by

$$\begin{aligned}\widehat{\mathcal{R}}_T(f, \rho) &= -2\rho + \frac{1}{m} \sum_{i=1}^m \ell(\rho - y_i f(x_i)), \\ \widehat{\mathcal{R}}_{T, \lambda}(f, \rho) &= -2\rho + \frac{1}{m} \sum_{i=1}^m \ell(\rho - y_i f(x_i)) + \theta(\|f\|_{\mathcal{H}}^2 \leq \lambda^2).\end{aligned}$$

The subscript T is dropped if it is clear from the context.

For the observed training samples $T_1 = \{(x_i^{(1)}, y_i^{(1)}) : i = 1, \dots, m_1\}$, clearly the problem (28) is identical to the minimization of $\widehat{\mathcal{R}}_{T_1, \lambda}(f, \rho)$. We define \widehat{f}, \widehat{b} and $\widehat{\rho}$ as an optimal solution of

$$\min_{f, b, \rho} \widehat{\mathcal{R}}_{T_1, \lambda_{m_1}}(f + b, \rho), \quad f \in \mathcal{H}, b \in \mathbb{R}, \rho \in \mathbb{R}, \quad (29)$$

where the regularization parameter λ_{m_1} may depend on the sample size. For the index sets M_p and M_n in Figure 5, we define $m_p = |M_p|$ and $m_n = |M_n|$.

We introduce the following assumptions.

Assumption 1 (universal kernel). *The input space \mathcal{X} is a compact metric space. The kernel function $k : \mathcal{X}^2 \rightarrow \mathbb{R}$ is continuous, and satisfies*

$$\sup_{x \in \mathcal{X}} \sqrt{k(x, x)} \leq K < \infty,$$

where K is a positive constant. In addition, k is universal, i.e., the RKHS associated with k is dense in the set of all continuous functions on \mathcal{X} with respect to the supremum norm [27, Definition 4.52].

Assumption 2 (non-deterministic assumption). *For the probability distribution of training samples, there exists a positive constant $\varepsilon > 0$ such that*

$$P(\{x \in \mathcal{X} : \varepsilon \leq P(+1|x) \leq 1 - \varepsilon\}) > 0$$

holds, where $P(y|x)$ is the conditional probability of the label y for given input x .

Assumption 3 (basic assumptions on the loss function). *The loss function $\ell : \mathbb{R} \rightarrow \mathbb{R}$ satisfies the following conditions.*

1. ℓ is a non-decreasing, convex function, and satisfies the non-negativity condition, i.e., $\ell(z) \geq 0$ for all $z \in \mathbb{R}$.
2. Let $\partial\ell(z)$ be the subdifferential of the loss function ℓ at $z \in \mathbb{R}$ [21, Chap. 23]. Then, the equality $\lim_{z \rightarrow \infty} \partial\ell(z) = \infty$ holds, i.e., for any $M > 0$, there exists z_0 such that for all $z \geq z_0$ and all $g \in \partial\ell(z)$, the inequality $g \geq M$ holds.

Note that the second condition in Assumption 3 assures that ℓ is not constant function and that $\lim_{z \rightarrow \infty} \ell(z) = \infty$ holds.

Assumption 4 (modified classification-calibrated loss).

1. $\ell(z)$ is first order differentiable for $z \geq -\ell(0)/2$, and $\ell'(z) > 0$ holds for $z \geq -\ell(0)/2$, where ℓ' is the derivative of ℓ .
2. Let $\psi(\theta, \rho)$ be the function defined as

$$\psi(\theta, \rho) = \ell(\rho) - \inf_{z \in \mathbb{R}} \left\{ \frac{1+\theta}{2} \ell(\rho - z) + \frac{1-\theta}{2} \ell(\rho + z) \right\}, \quad 0 \leq \theta \leq 1, \quad \rho \in \mathbb{R}.$$

There exist a function $\tilde{\psi}(\theta)$ and a positive real $\varepsilon > 0$ such that the following conditions are satisfied:

- (a) $\tilde{\psi}(0) = 0$ and $\tilde{\psi}(\theta) > 0$ for $0 < \theta \leq \varepsilon$.
- (b) $\tilde{\psi}(\theta)$ is a continuous and strictly increasing function on the interval $[0, \varepsilon]$.
- (c) The inequality $\tilde{\psi}(\theta) \leq \inf_{\rho \geq -\ell(0)/2} \psi(\theta, \rho)$ holds for $0 \leq \theta \leq \varepsilon$.

Later on, we shall give some sufficient conditions for existence of the function $\tilde{\psi}$ in Assumption 4.

We prove that there is no duality gap between (24) and (28). The proof of the following lemma is given in Appendix A.

Lemma 1. *Suppose that both M_p and M_n in Figure 5 are non-empty, i.e., m_p and m_n are positive numbers. Under Assumption 1 and 3, there exists an optimal solution for (28). Moreover, the dual problem of (28) yields the problem (24) with the uncertainty set (27).*

In the following, we prove the convergence of the error rate to the Bayes risk \mathcal{E}^* . The proof consists of two parts. In Section 6.2, we prove that the expected loss for the estimated decision function, $\mathcal{R}(\hat{f} + \hat{b}, \hat{\rho})$, converges to the infimum of the expected loss \mathcal{R}^* , where \hat{f}, \hat{b} and $\hat{\rho}$ are optimal solutions of (29). Here, we apply the mathematical tools developed by [26]. In Section 6.3, we prove the convergence of the error rate $\mathcal{E}(\hat{f} + \hat{b})$ to the Bayes risk \mathcal{E}^* , where \hat{b} is an optimal solution of (26). In the proof, the concept of the classification-calibrated loss [2] plays an important role.

6.2 Convergence to Optimal Expected Loss

In this section, we prove that $\mathcal{R}(\hat{f} + \hat{b}, \hat{\rho})$ converges to \mathcal{R}^* . Following lemmas show the relation between the expected loss and the regularized the expected loss. Proofs are shown in Appendix B.

Lemma 2. *Under Assumption 2 and Assumption 3, we have $\mathcal{R}^* > -\infty$.*

Lemma 3. *Under Assumption 1, 2 and 3, we have*

$$\lim_{\lambda \rightarrow \infty} \inf \{ \mathcal{R}_\lambda(f, \rho) : f \in \mathcal{H}, \rho \in \mathbb{R} \} = \mathcal{R}^*. \quad (30)$$

We derive an upper bound on the norm of the optimal solution in (29). The proof is deferred to Appendix B.

Lemma 4. *Under Assumption 1, 2 and 3, there are positive constants c and C and a natural number M such that the optimal solution of (29) satisfies*

$$\|\hat{f}\|_{\mathcal{H}} \leq \lambda_{m_1}, \quad |\hat{b}| \leq C\lambda_{m_1}, \quad |\hat{\rho}| \leq C\lambda_{m_1} \quad (31)$$

with the probability greater than $1 - e^{-cm_1}$ for $m_1 \geq M$.

Let us define the covering number for a metric space.

Definition 1 (covering number). *For a metric space \mathcal{G} , the covering number of \mathcal{G} is defined as*

$$\mathcal{N}(\mathcal{G}, \varepsilon) = \min \left\{ n \in \mathbb{N} : g_1, \dots, g_n \in \mathcal{G} \text{ such that } \mathcal{G} \subset \bigcup_{i=1}^n B(g_i, \varepsilon) \right\},$$

where $B(g, \varepsilon)$ denotes the closed ball with center g and radius ε .

According to Lemma 4, the optimal solution, \hat{f} , \hat{b} and $\hat{\rho}$, is included in the set

$$\mathcal{G}_{m_1} = \{(f, b, \rho) \in \mathcal{H} \times \mathbb{R}^2 : \|f\|_{\mathcal{H}} \leq \lambda_{m_1}, |b| \leq C\lambda_{m_1}, |\rho| \leq C\lambda_{m_1}\}$$

with high probability. Suppose that the norm $\|f\|_{\infty} + |b| + |\rho|$ is introduced on \mathcal{G}_{m_1} . We define the function

$$L(x, y; f, b, \rho) = -2\rho + \ell(\rho - y(f(x) + b)),$$

and the function set

$$\mathcal{L}_{m_1} = \{L(x, y; f, b, \rho) : (f, b, \rho) \in \mathcal{G}_{m_1}\}.$$

The supremum norm is defined on \mathcal{L}_{m_1} . The expected loss and the empirical loss, $\mathcal{R}(f + b, \rho)$ and $\hat{\mathcal{R}}_{T_1}(f + b, \rho)$, are represented as the expectation of $L(x, y; f, b, \rho)$ with respect to the population distribution and the empirical distribution, respectively. Since $\ell : \mathbb{R} \rightarrow \mathbb{R}$ is a finite-valued convex function, ℓ is locally Lipschitz continuous. Then, for any sample size m_1 , there exists a constant κ_{m_1} depending on m_1 such that

$$|\ell(z) - \ell(z')| \leq \kappa_{m_1} |z - z'| \quad (32)$$

holds for all z and z' satisfying $|z|, |z'| \leq (K + 2C)\lambda_{m_1}$. Then, for any $(f, b, \rho), (f', b', \rho') \in \mathcal{G}_{m_1}$, we have

$$\begin{aligned} |L(x, y; f, b, \rho) - L(x, y; f', b', \rho')| &\leq 2|\rho - \rho'| + \kappa_{m_1}(|\rho - \rho'| + |b - b'| + \|f - f'\|_\infty) \\ &\leq (2 + \kappa_{m_1})(|\rho - \rho'| + |b - b'| + \|f - f'\|_\infty) \end{aligned}$$

The covering number of \mathcal{L}_{m_1} is evaluated by using that of \mathcal{G}_{m_1} as follows:

$$\mathcal{N}(\mathcal{L}_{m_1}, \varepsilon) \leq \mathcal{N}(\mathcal{G}_{m_1}, \frac{\varepsilon}{2 + \kappa_{m_1}}). \quad (33)$$

Let the metric space \mathcal{F}_{m_1} be

$$\mathcal{F}_{m_1} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda_{m_1}\}$$

with the supremum norm, then, we also have

$$\mathcal{N}\left(\mathcal{G}_{m_1}, \frac{\varepsilon}{2 + \kappa_{m_1}}\right) \leq \mathcal{N}\left(\mathcal{F}_{m_1}, \frac{\varepsilon}{3(2 + \kappa_{m_1})}\right) \left(\frac{6C\lambda_{m_1}(2 + \kappa_{m_1})}{\varepsilon}\right)^2. \quad (34)$$

An upper bound of the covering number of \mathcal{F}_{m_1} is given by [10] and [31].

We prove the uniform convergence of $\widehat{\mathcal{R}}(f + b, \rho)$. The proof is deferred to Appendix B.

Lemma 5. *Let b_{m_1} be*

$$b_{m_1} = 4C\lambda_{m_1} + \ell((K + 2C)\lambda_{m_1})$$

in which C is the positive constant defined in Lemma 4. Under Assumption 1 and 3, the inequality

$$\begin{aligned} &P\left(\sup_{(f, b, \rho) \in \mathcal{G}_{m_1}} |\widehat{\mathcal{R}}(f + b, \rho) - \mathcal{R}(f + b, \rho)| \geq \varepsilon\right) \\ &\leq 2\mathcal{N}(\mathcal{L}_{m_1}, \varepsilon/3) \exp\left\{-\frac{2m_1\varepsilon^2}{9b_{m_1}^2}\right\} \end{aligned} \quad (35)$$

$$\leq 2\mathcal{N}\left(\mathcal{F}_{m_1}, \frac{\varepsilon}{9(2 + \kappa_{m_1})}\right) \left(\frac{18C\lambda_{m_1}(2 + \kappa_{m_1})}{\varepsilon}\right)^2 \exp\left\{-\frac{2m_1\varepsilon^2}{9b_{m_1}^2}\right\} \quad (36)$$

holds, where κ_{m_1} is the Lipschitz constant defined by (32).

We present the main theorem of this section. The proof is given in Appendix C.

Theorem 2. *Suppose that $\lim_{m_1 \rightarrow \infty} \lambda_{m_1} = \infty$ holds. Suppose that Assumption 1, 2 and 3 hold. Moreover we assume that (36) converges to zero for any $\varepsilon > 0$, when the sample size m_1 tends to infinity. Then, $\mathcal{R}(\widehat{f} + \widehat{b}, \widehat{\rho})$ converges to \mathcal{R}^* in probability in the large sample limit of the dataset T_1 .*

We show the order of λ_{m_1} admitting the assumption in Theorem 2.

Example 7. Suppose that $\mathcal{X} = [0, 1]^n \subset \mathbb{R}^n$ and the Gaussian kernel is used. According to [31], we have

$$\log \mathcal{N}\left(\mathcal{F}_{m_1}, \frac{\varepsilon}{9(2 + \kappa_{m_1})}\right) = O\left(\left(\log \frac{\lambda_{m_1}}{\frac{\varepsilon}{9(2 + \kappa_{m_1})}}\right)^{n+1}\right) = O((\log(\lambda_{m_1} \kappa_{m_1}))^{n+1}).$$

For any $\varepsilon > 0$, (36) is bounded above by

$$\exp\left\{O\left(-\frac{m_1}{b_{m_1}^2} + (\log(\lambda_{m_1} \kappa_{m_1}))^{n+1}\right)\right\}.$$

For the truncated quadratic loss, we have

$$\begin{aligned}\kappa_{m_1} &\leq 2((K + 2C)\lambda_{m_1} + 1) = O(\lambda_{m_1}), \\ b_{m_1} &\leq 4C\lambda_{m_1} + ((K + 2C)\lambda_{m_1} + 1)^2 = O(\lambda_{m_1}^2).\end{aligned}$$

Let us define $\lambda_{m_1} = m_1^\alpha$ with $0 < \alpha < 1/4$. Then, for any $\varepsilon > 0$, (36) converges to zero when m_1 tends to infinity. In the same way, for the exponential loss we obtain

$$\kappa_{m_1} = O(e^{(K+2C)\lambda_{m_1}}), \quad b_{m_1} = O(e^{(K+2C)\lambda_{m_1}}).$$

Hence, $\lambda_{m_1} = (\log m_1)^\alpha$ with $0 < \alpha < 1$ assures the convergence of (36).

6.3 Convergence to Bayes Risk

We study the error rate of the estimated classifier. Let us define \hat{f}, \hat{b} and $\hat{\rho}$ be a minimizer of $\mathcal{R}_{T_1, \lambda_{m_1}}(f + b, \rho)$. In the proposed learning algorithm in Figure 5, the estimated bias term \hat{b} is replaced with \tilde{b} which is an optimal solution of $\min_{b \in \mathbb{R}} \hat{\mathcal{E}}_{T_2}(\hat{f} + b)$. We prove that the expected 0-1 loss $\mathcal{E}(\hat{f} + \tilde{b})$ converges to the Bayes risk \mathcal{E}^* , when the sample sizes of T_1 and T_2 tend to infinity. The proof is shown in Appendix D.

Theorem 3. Suppose that $\mathcal{R}(\hat{f} + \hat{b}, \hat{\rho})$ converges to \mathcal{R}^* in probability, when the sample size of T_1 , i.e., m_1 , tends to infinity. For the RKHS \mathcal{H} and the loss function ℓ , we assume Assumption 1, 3 and 4. Then, $\mathcal{E}(\hat{f} + \tilde{b})$ converges to \mathcal{E}^* in probability, when the sample sizes of T_1 and T_2 tend to infinity.

As a result, we find that the prediction error rate of $\hat{f} + \tilde{b}$ converges to the Bayes risk under Assumption 1, 2, 3 and 4.

We present some sufficient conditions for existence of the function $\tilde{\psi}$ in Assumption 4. The proof of the following lemma is shown in Appendix E.

Lemma 6. Suppose that the first condition in Assumption 3 and the first condition in Assumption 4 hold. In addition, suppose that ℓ is first-order continuously differentiable on \mathbb{R} . Let d be $d = \sup\{z \in \mathbb{R} : \ell'(z) = 0\}$, where ℓ' is the derivative of ℓ . When $\ell'(z) > 0$ holds for all $z \in \mathbb{R}$, we define $d = -\infty$. We assume the following conditions:

1. $d < -\ell(0)/2$.
2. $\ell(z)$ is second-order continuously differentiable on the open interval (d, ∞) .

3. $\ell''(z) > 0$ holds on (d, ∞) .

4. $1/\ell'(z)$ is convex on (d, ∞) .

Then, for any $\theta \in [0, 1]$, the function $\psi(\theta, \rho)$ is non-decreasing as the function of ρ for $\rho \geq -\ell(0)/2$.

When the condition in Lemma 6 is satisfied, we can choose $\psi(\theta, -\ell(0)/2)$ as $\tilde{\psi}(\theta)$ for $0 \leq \theta \leq 1$, since $\psi(\theta, -\ell(0)/2)$ is classification-calibrated under the first condition in Assumption 4.

We give another sufficient condition for existence of the function $\tilde{\psi}$ in Assumption 4. The proof of the following lemma is shown in Appendix E.

Lemma 7. Suppose that the first condition in Assumption 3 and the first condition in Assumption 4 hold. Let $d = \sup\{z \in \mathbb{R} : \partial\ell(z) = \{0\}\}$. When $0 \notin \partial\ell'(z)$ holds for all $z \in \mathbb{R}$, we define $d = -\infty$. Suppose that the inequality $-\ell(0)/2 > d$ holds. For $\rho \geq -\ell(0)/2$ and $z \geq 0$, we define $\xi(z, \rho)$ by

$$\xi(z, \rho) = \begin{cases} \frac{\ell(\rho + z) + \ell(\rho - z) - 2\ell(\rho)}{z\ell'(\rho)}, & z > 0, \\ 0, & z = 0. \end{cases}$$

Suppose that there exists a function $\bar{\xi}(z)$ for $z \geq 0$ such that the following conditions hold:

1. $\bar{\xi}(z)$ is continuous and strictly increasing on $z \geq 0$, and satisfies $\bar{\xi}(0) = 0$ and $\lim_{z \rightarrow \infty} \bar{\xi}(z) > 1$.
2. $\sup_{\rho \geq -\ell(0)/2} \xi(z, \rho) \leq \bar{\xi}(z)$ holds.

Then, there exists a function $\tilde{\psi}$ defined in the second condition of Assumption 4.

Note that Lemma 7 does not require the second order differentiability of the loss function. We show some examples in which the existence of $\tilde{\psi}$ is confirmed from the above lemmas.

Example 8. For the truncated quadratic loss $\ell(z) = (\max\{z + 1, 0\})^2$, the first condition in Assumption 3 and the first condition in Assumption 4 hold. The inequality $-\ell(0)/2 = -1/2 > \sup\{z : \ell'(z) = 0\} = -1$ in the sufficient condition of Lemma 6 holds. For $z > -1$, it is easy to see that $\ell(z)$ is second-order differentiable and that $\ell''(z) > 0$ holds. In addition, for $z > -1$, $1/\ell'(z)$ is equal to $1/(2z + 2)$ which is convex on $(-1, \infty)$. Therefore, the function $\tilde{\psi}(\theta) = \psi(\theta, -1/2)$ satisfies the second condition in Assumption 4.

Example 9. For the exponential loss $\ell(z) = e^z$, we have $1/\ell'(z) = e^{-z}$. Hence, due to Lemma 6, $\psi(\theta, \rho)$ is non-decreasing in ρ . Indeed, we have $\psi(\theta, \rho) = (1 - \sqrt{1 - \theta^2})e^\rho$.

Example 10. In Example 6, we presented the uncertainty set with estimation errors. The uncertainty sets are defined based on the revised function $\bar{\ell}(z)$ in (22). Here, we use a similar function defined by

$$\bar{\ell}^*(\alpha) = \begin{cases} (|\alpha w - 1| + h)^2 - (1 + h)^2, & \alpha \geq 0, \\ \infty, & \alpha < 0, \end{cases} \quad (37)$$

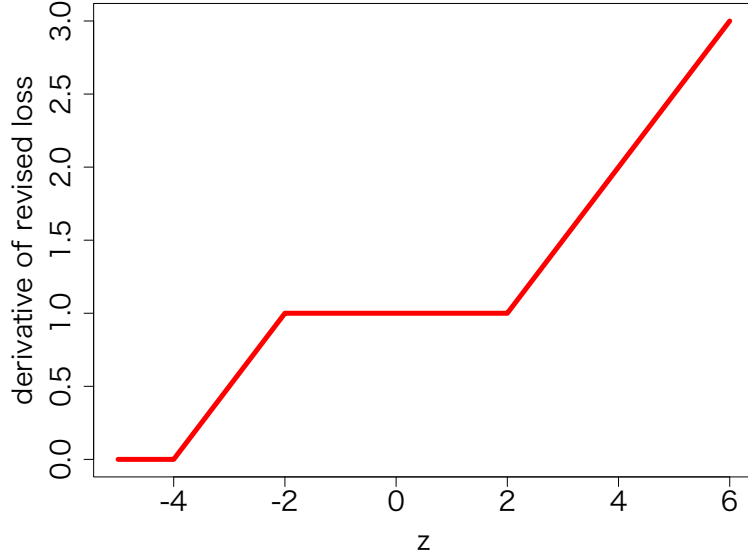


Figure 6: The derivative of the loss function corresponding to the revised uncertainty set with the estimation error.

for the construction of uncertainty sets. Here, w and h are positive constants, and we suppose $w > 1/2$. The corresponding loss function is given as $\bar{\ell}(z)$. Then we have $\bar{\ell}(z) = u(z/w)$ defined in (23). For $w > 1/2$, we can confirm that $\sup\{z : \bar{\ell}'(z) = 0\} < -\bar{\ell}(0)/2$ holds. Since $u(z)$ is not strictly convex, Lemma 6 does not work. Hence, we apply Lemma 7. A simple calculation yields that $\bar{\ell}'(-\bar{\ell}(0)/2) \geq (4w-1)/(4w^2) > 0$ for any $h \geq 0$. Note that $\bar{\ell}(z)$ is differentiable on \mathbb{R} . Thus, the monotonicity of $\bar{\ell}'$ for the convex function leads to

$$\xi(z, \rho) = \frac{1}{\bar{\ell}'(\rho)} \left(\frac{\bar{\ell}(\rho+z) - \bar{\ell}(\rho)}{z} - \frac{\bar{\ell}(\rho) - \bar{\ell}(\rho-z)}{z} \right) \leq \frac{\bar{\ell}'(\rho+z) - \bar{\ell}'(\rho-z)}{\bar{\ell}'(\rho)}.$$

Figure 6 depicts the derivative of $\bar{\ell}$ with $h = 1$ and $w = 1$. Since the derivative $\bar{\ell}'(z)$ is Lipschitz continuous and the Lipschitz constant is equal to $1/(2w)$, we have $\bar{\ell}'(\rho+z) - \bar{\ell}'(\rho-z) \leq z/w$. Therefore, the inequality

$$\sup_{\rho \geq -\bar{\ell}(0)/2} \xi(z, \rho) \leq \sup_{\rho \geq -\bar{\ell}(0)/2} \frac{z/w}{\bar{\ell}'(\rho)} = \frac{z/w}{\bar{\ell}'(-\bar{\ell}(0)/2)} \leq \frac{4w}{4w-1} z \leq 2z$$

holds. We see that $\bar{\xi}(z) = 2z$ satisfies the sufficient condition of Lemma 7. The inequality

$$\bar{\ell}'(-\bar{\ell}(0)/2) \frac{\theta}{2} \bar{\xi}^{-1}\left(\frac{\theta}{2}\right) \geq \frac{4w-1}{32w^2} \theta^2$$

ensures that $\tilde{\psi}(\theta) = \frac{4w-1}{32w^2} \theta^2$ is a valid choice. Therefore, the loss function corresponding to the revised uncertainty set in Example 6 satisfies the sufficient conditions for the Bayes risk consistency.

7 Experiments

We compare the statistical properties of the proposed learning algorithm to the other learning methods. As proved in Section 6, the kernel-based learning algorithm in Figure 5 has the statistical consistency under some assumptions, while MPM and MM-MPM do not have the statistical consistency in general. The main purpose of the numerical study is to compare our method to MPM and its variants.

We compare the kernel-based learning algorithms using the Gaussian kernel. So far, many works have been devoted to compare the linear models and the kernel-based models. The conclusion is that the linear model outperforms the kernel-based model when the decision boundary is well approximated by the linear model. Otherwise, the linear model has the approximation bias, and the kernel-based estimators with a nice regularization outperform the linear models in general. Hence, we focus on the kernel-based estimators. In our experiments, the following methods were examined to the synthetic data and the standard benchmark datasets: C -SVM, MPM, unbiased MPM, and the kernel variant of the proposed method presented in Figure 2. For simplicity, the function part $f \in \mathcal{H}$ and the bias term $b \in \mathbb{R}$ are estimated based on all training samples, though in the learning algorithm in Figure 5, the dataset is decomposed into two subsets in order to ensure the statistical consistency. In the unbiased MPM, the bias term b in the model is estimated by minimizing the training error rate after estimating the function part, $\hat{f} \in \mathcal{H}$. Clearly, the unbiased estimator will outperform the original MPM, when the probability of the class label is heavily unbalanced. In the proposed method, we apply the uncertainty set defined from the loss function $u(z)$ defined in (23). This is the revised uncertainty set of the ellipsoidal uncertainty set with the estimation error. The parameter in the function $u(z)$ of (23) is set to $h = 0$ or $h = 1$. The kernel parameter and the regularization parameter are estimated by 5-fold cross validation. We use the test error for the evaluation of the prediction accuracy.

7.1 Synthetic data

Suppose that the input points \mathbf{x} conditioned on the positive label are generated by the two dimensional normal distribution with the mean $\boldsymbol{\mu}_p = (0, 0)^T$ and the covariance matrix $\Sigma_p = I$, where I is the identity matrix. In the same way, the conditional distribution of input points with the negative label is defined as the normal distribution with $\boldsymbol{\mu}_n = (1, 1)^T$ and the covariance matrix $\Sigma_n = R^T \text{diag}(0.5^2, 1.5^2) R$, where R is the $\pi/3$ radian counterclockwise rotation matrix. The label probability is defined by $P(Y = +1) = 0.2$ or 0.5 . The size of training samples is $m = 400$.

Table 3 shows the test error of the estimators: C -SVM, MPM, unbiased MPM, learning with the loss function (23) with $h = 0$ or $h = 1$. We notice that, under the unbalanced samples, i.e., the case of $P(Y = +1) = 0.2$, the MPM has the estimation bias. On the setup of the balanced data, MPM is slightly better than the other methods. All the learning algorithm except MPM are comparable to each other. The difference of the parameter h in the loss function (23) is not significant in this experiment.

7.2 Benchmark data

In this section, we use thirteen artificial and real world datasets from the UCI, DELVE, and STATLOG benchmark repositories: `banana`, `breast-cancer`, `diabetes`, `german`, `heart`, `image`, `ringnorm`, `flare-solar`, `splice`, `thyroid`, `titanic`, `twonorm`, `waveform`. All datasets are

Table 1: Test error (%) of each learning method is presented with the standard deviation. We compared C -SVM, MPM, unbiased MPM, learning method with the loss function (23) with $h = 0$ and $h = 1$.

$P(Y=+1)$	C -SVM	MPM	unbiased MPM	$h = 0$	$h = 1$
0.2	15.8 ± 1.1	26.0 ± 2.2	16.5 ± 1.2	15.9 ± 1.1	16.0 ± 1.2
0.5	25.2 ± 1.1	25.1 ± 1.0	25.5 ± 1.3	25.5 ± 1.4	25.4 ± 1.1

provided as IDA benchmark repository. See [20] and [19] for details of datasets. The properties of each dataset are shown in Table 2, where “dim”, “ $P(Y = +1)$ ”, “#train”, “#test” and “rep.” denote the input dimension, the ratio of the positive labels in training samples, the size of training set, the size of test set, and the number of replication of learning to evaluate the average performance, respectively.

In the experiment, especially we compare unbiased MPM and our method using the loss function (23) with $h = 0$. The uncertainty set of unbiased MPM is ellipsoid defined by the estimated covariance matrix. The corresponding loss function of the form of (8) does not exist, since the convex-hull of the input points is not taken into account. In our method using the loss function (23) with $h = 0$, the uncertainty set is the intersection of the same ellipsoid as unbiased MPM and the convex-hull of the input vectors. That is, the revision of the ellipsoidal uncertainty set in unbiased MPM leads to the uncertainty set of our algorithm. We use the t -test to detect the difference of test errors of these two learning algorithms.

Table 3 shows test errors (%) for benchmark datasets with the standard deviation. We show the results of C -SVM, MPM, unbiased MPM, learning method with the loss function (23) with $h = 0$ and $h = 1$. In the column of the unbiased MPM and our method with $h = 0$, the bold face letters indicates that the test error is smaller compared to the opponent at the significance level 1%. Overall, C -SVM performs better than the others. the learning method with the loss function (23) with $h = 1$ is comparable to C -SVM except **breast-cancer**, **flare-solar** and **titanic**. Note that the loss function (23) with $h = 1$ is similar to the hinge loss around zero. Hence, it is clear that the results of our method with $h = 1$ is close to the results of C -SVM. The results of t -test indicates that, comparing to unbiased MPM, our method using the loss function (23) with $h = 0$ achieves the smaller test errors. In both algorithms, the same estimator is used for the bias term in the decision function. Hence, the result implies that our method is superior to unbiased MPM in the estimation of the function part $f \in \mathcal{H}$ in the decision function. In the dataset **flare-solar** and **titanic**, unbiased MPM is superior to our method with $h = 0$. This is because there are many duplications in covariates of these datasets. Indeed, in 666 training samples of **flare-solar**, there are only 76 different input points, and **titanic** has only 11 different input points out of 150 training samples. In the other datasets, the variety of the covariates is almost equal to the size of the training samples. In our method, the uncertainty set for such data does not capture the distribution of the input points appropriately. We notice that the revision of the uncertainty set will be useful to achieve high prediction accuracy in comparison to (unbiased) MPM, as long as the covariate does not have many duplications.

8 Conclusion

In this paper, we studied the relation between the loss function approach and the uncertainty set approach in binary classification problems. We showed that these two approaches are con-

Table 2: The properties of each data sets are shown, where “dim”, “ $P(Y = +1)$ ”, “#train”, “#test” and “rep.” denote the input dimension, the ratio of the positive label in training samples, the size of training set, the size of test set, and the number of replication of learning, respectively.

dataset	dim	$P(Y = +1)$	#train	#test	rep.
banana	2	0.454	400	4900	100
breast-cancer	9	0.294	200	77	100
diabetis	8	0.350	468	300	100
flare-solar	9	0.552	666	400	100
german	20	0.301	700	300	100
heart	13	0.445	170	100	100
image	18	0.574	1300	1010	20
ringnorm	20	0.497	400	7000	100
splice	60	0.483	1000	2175	20
thyroid	5	0.305	140	75	85
titanic	3	0.322	150	2051	100
twonorm	20	0.505	400	7000	100
waveform	21	0.331	400	4600	100

Table 3: Test errors (%) for benchmark datasets are presented with the standard deviation. We compared C -SVM, MPM, unbiased MPM, learning method with the loss function (23) with $h = 0$ and $h = 1$. We conduct t -test to compare the unbiased MPM and the learning method using the loss function (23) with $h = 0$. The bold face letters indicates that the test error is smaller compared to the opponent at the significance level 1%.

dataset	C -SVM	MPM	unbiased MPM	$h = 0$	$h = 1$
banana	10.7 ± 0.6	11.4 ± 0.9	11.4 ± 0.9	11.1 ± 0.9	10.9 ± 0.7
breast-cancer	26.9 ± 4.8	35.0 ± 4.9	34.0 ± 4.8	28.1 ± 5.0	28.1 ± 4.5
diabetis	23.9 ± 2.1	28.8 ± 2.4	28.3 ± 2.5	24.3 ± 1.9	24.2 ± 2.1
flare-solar	33.7 ± 2.2	34.9 ± 1.7	35.7 ± 1.9	36.8 ± 3.1	36.8 ± 2.9
german	23.8 ± 2.3	29.2 ± 2.4	28.2 ± 2.7	23.5 ± 2.3	23.6 ± 2.4
heart	16.7 ± 3.5	25.6 ± 4.2	25.7 ± 4.0	17.3 ± 3.7	17.2 ± 3.5
image	3.3 ± 0.7	3.2 ± 0.7	3.2 ± 0.7	3.4 ± 0.6	3.3 ± 0.5
ringnorm	1.7 ± 0.3	3.2 ± 0.4	2.8 ± 0.5	1.7 ± 0.3	1.6 ± 0.2
splice	11.1 ± 0.7	12.3 ± 1.7	11.7 ± 0.8	11.3 ± 0.7	11.1 ± 0.8
thyroid	5.3 ± 2.1	6.3 ± 3.1	6.2 ± 3.7	5.6 ± 2.4	5.4 ± 2.2
titanic	22.4 ± 0.8	24.1 ± 2.2	22.4 ± 1.2	23.5 ± 1.6	23.7 ± 3.4
twonorm	2.6 ± 0.3	4.5 ± 0.7	4.4 ± 0.6	2.6 ± 0.3	2.6 ± 0.4
waveform	10.2 ± 0.7	13.0 ± 0.9	12.7 ± 0.8	10.2 ± 0.6	10.1 ± 0.7

nected to each other by the conjugate property based on the Legendre transformation. Given a loss function, there exists a corresponding parametrized uncertainty set. In general, however, uncertainty set does not correspond to the empirical loss function. We presented a way of revising the uncertainty set such that there exists an empirical loss function. Then, we proposed a modified maximum-margin algorithm based on the parametrized uncertainty set. We proved

the statistical consistency of the learning algorithm. Numerical experiments showed that the revision of the uncertainty set often improves the prediction accuracy of the classifier.

In our proof of the statistical consistency, the hinge loss used in ν -SVM is excluded. [25] proved the statistical consistency of ν -SVM with a nice choice of the regularization parameter. We are currently investigating the relaxation of the assumptions of our theoretical result so as to include the hinge loss function and other popular loss functions such as the logistic loss. As for the statistical modeling, the relation between the loss function approach and the uncertainty set approach can be a useful tool. In optimization and control theory, the modeling based on the uncertainty set is frequently applied to the real-world data; see the modeling in robust optimization and related works [3]. We believe that the learning algorithm with the revision of the uncertainty set can bridge a gap between statistical modeling based on some intuition and nice statistical properties of the estimated classifiers.

Acknowledgments

TK was partially supported by Grant-in-Aid for Young Scientists (20700251). AT was partially supported by Grant-in-Aid for Young Scientists (23710174). TS was partially supported by MEXT Kakenhi 22700289 and the Aihara Project, the FIRST program from JSPS, initiated by CSTP.

A Proof of Lemma 1

First, we prove the existence of an optimal solution. According to the standard argument on the kernel estimator, we can restrict the function part f to be the form of

$$f(x) = \sum_{j=1}^{m_1} \alpha_j k(x, x_j^{(1)}).$$

Then, the problem is reduced to the finite-dimensional problem,

$$\begin{aligned} \min_{\alpha, b, \rho} & -2\rho + \frac{1}{m_1} \sum_{i=1}^{m_1} \ell(\rho - y_i^{(1)} (\sum_{j=1}^{m_1} \alpha_j k(x_i^{(1)}, x_j^{(1)}) + b)) \\ \text{subject to} & \sum_{i,j=1}^{m_1} \alpha_i \alpha_j k(x_i^{(1)}, x_j^{(1)}) \leq \lambda^2. \end{aligned} \tag{38}$$

Let $\zeta_0(\alpha, b, \rho)$ be the objective function of (38). Let us define \mathcal{S} be the linear subspace in \mathbb{R}^{m_1} spanned by the column vectors of the gram matrix $(k(x_i^{(1)}, x_j^{(1)}))_{i,j=1}^{m_1}$. We can impose the constraint $\alpha = (\alpha_1, \dots, \alpha_{m_1}) \in \mathcal{S}$, since the orthogonal complement of \mathcal{S} does not affect the objective and the constraint in (38). We see that Assumption 1 and the reproducing property yield the inequality $\|y_i^{(1)} \sum_{j=1}^{m_1} \alpha_j k(\cdot, x_j^{(1)})\|_\infty \leq K\lambda$. Due to this inequality and the assumptions on the function ℓ , the objective function $\zeta_0(\alpha, b, \rho)$ is bounded below by

$$\zeta_1(b, \rho) = -2\rho + \frac{m_p}{m_1} \ell(\rho - b - K\lambda) + \frac{m_n}{m_1} \ell(\rho + b - K\lambda). \tag{39}$$

Hence, for any real number c , the inclusion relation

$$\begin{aligned} & \left\{ (\alpha, b, \rho) \in \mathbb{R}^{m_1+2} : \zeta_0(\alpha, b, \rho) \leq c, \sum_{i,j=1}^{m_1} \alpha_i \alpha_j k(x_i^{(1)}, x_j^{(1)}) \leq \lambda^2, \alpha \in \mathcal{S} \right\} \\ & \subset \left\{ (\alpha, b, \rho) \in \mathbb{R}^{m_1+2} : \zeta_1(b, \rho) \leq c, \sum_{i,j=1}^{m_1} \alpha_i \alpha_j k(x_i^{(1)}, x_j^{(1)}) \leq \lambda^2, \alpha \in \mathcal{S} \right\} \end{aligned} \quad (40)$$

holds. Note that the vector α satisfying $\sum_{i,j=1}^{m_1} \alpha_i \alpha_j k(x_i^{(1)}, x_j^{(1)}) \leq \lambda^2$ and $\alpha \in \mathcal{S}$ is restricted to a compact subset in \mathbb{R}^{m_1} . We shall prove that the subset (40) is compact, if they are not empty. We see that the two sets above are closed subsets, since both ζ_0 and ζ_1 are continuous. By the variable change from (b, ρ) to $(u_1, u_2) = (\rho - b, \rho + b)$, $\zeta_1(b, \rho)$ is transformed to the convex function $\zeta_2(u_1, u_2)$ defined by

$$\zeta_2(u_1, u_2) = -u_1 + \frac{m_p}{m_1} \ell(u_1 - K\lambda) - u_2 + \frac{m_n}{m_1} \ell(u_2 - K\lambda).$$

The subgradient of $\ell(z)$ diverges to infinity, when z tends to infinity. In addition, $\ell(z)$ is a non-decreasing and non-negative function. Then, we have

$$\lim_{|u_1| \rightarrow \infty} -u_1 + \frac{m_p}{m_1} \ell(u_1 - K\lambda) = \infty.$$

The same limit holds for $-u_2 + \frac{m_n}{m_1} \ell(u_2 - K\lambda)$. Hence, the level set of $\zeta_2(u_1, u_2)$ is closed and bounded, i.e., compact. As a result, the level set of $\zeta_1(b, \rho)$ is also compact. Therefore, the subset (40) is also compact in \mathbb{R}^{m_1+2} . This implies that (38) has an optimal solution.

Next, we prove the duality between (29) and (24). Since (38) has an optimal solution, the problem with the slack variables $\xi_i, i = 1, \dots, m_1$,

$$\begin{aligned} & \min_{\alpha, b, \rho, \xi} -2\rho + \frac{1}{m_1} \sum_{i=1}^{m_1} \ell(\xi_i) \\ & \text{subject to } \sum_{i,j=1}^{m_1} \alpha_i \alpha_j k(x_i^{(1)}, x_j^{(1)}) \leq \lambda^2, \\ & \rho - y_i^{(1)} \left(\sum_{j=1}^{m_1} \alpha_j k(x_i^{(1)}, x_j^{(1)}) + b \right) \leq \xi_i, i = 1, \dots, m_1. \end{aligned}$$

also has an optimal solution and the finite optimal value. In addition, the above problem clearly satisfies the Slater condition [7, Assumption 6.2.4]. Indeed, at the feasible solution, $\alpha = \mathbf{0}, b = 0, \rho = 0$ and $\xi_i = 1, i = 1, \dots, m_1$, the constraint inequalities are all inactive for positive λ . Hence, Proposition 6.4.3 in [7] ensures that the min-max theorem holds, i.e., there is no duality gap. Then, in the same way as (9), we obtain (24) with the uncertainty set (27) as the dual problem of (29).

B Proofs of Lemmas in Section 6.2

We show proofs of lemmas in Section 6.2.

B.1 Proof of Lemma 2

Let $S \subset \mathcal{X}$ be the subset $S = \{x \in \mathcal{X} : \varepsilon \leq P(+1|x) \leq 1 - \varepsilon\}$, then we have $P(S) > 0$. Due to the non-negativity of the loss function ℓ , we have

$$\begin{aligned}\mathcal{R}(f, \rho) &\geq -2\rho + \int_S \left\{ P(+1|x)\ell(\rho - f(x)) + P(-1|x)\ell(\rho + f(x)) \right\} P(dx) \\ &= \int_S \left\{ -\frac{2}{P(S)}\rho + P(+1|x)\ell(\rho - f(x)) + P(-1|x)\ell(\rho + f(x)) \right\} P(dx).\end{aligned}$$

For given η satisfying $\varepsilon \leq \eta \leq 1 - \varepsilon$, we define the function $\xi(f, \rho)$ by

$$\xi(f, \rho) = -\frac{2}{P(S)}\rho + \eta\ell(\rho - f) + (1 - \eta)\ell(\rho + f), \quad f, \rho \in \mathbb{R}.$$

We derive a lower bound $\inf\{\xi(f, \rho) : f, \rho \in \mathbb{R}\}$. Since $\ell(z)$ is a finite-valued convex function on \mathbb{R} , the subdifferential $\partial\xi(f, \rho) \subset \mathbb{R}^2$ is given as

$$\partial\xi(f, \rho) = \left\{ (0, -\frac{2}{P(S)})^T + u\eta(-1, 1)^T + v(1 - \eta)(1, 1)^T : u \in \partial\ell(\rho - f), v \in \partial\ell(\rho + f) \right\}.$$

Formulas of the subdifferential are presented in Theorem 23.8 and Theorem 23.9 of [21]. We prove that there exist f^* and ρ^* such that $(0, 0)^T \in \partial\xi(f^*, \rho^*)$ holds. Since the second condition in Assumption 3 holds for the convex function ℓ , the union $\cup_{z \in \mathbb{R}} \partial\ell(z)$ includes all the positive real numbers. Hence, there exist z_1 and z_2 satisfying $\frac{1}{\eta P(S)} \in \partial\ell(z_1)$ and $\frac{1}{(1-\eta)P(S)} \in \partial\ell(z_2)$. Then, for $f^* = (z_2 - z_1)/2$, $\rho^* = (z_1 + z_2)/2$, the null vector is an element of $\partial\xi(f^*, \rho^*)$. Since $\xi(f, \rho)$ is convex in (f, ρ) , the minimum value of $\xi(f, \rho)$ is attained at (f^*, ρ^*) . Define z_{up} as a real number satisfying

$$g > \frac{1}{\varepsilon P(S)}, \quad \forall g \in \partial\ell(z_{\text{up}}).$$

Since $\varepsilon \leq \eta \leq 1 - \varepsilon$ is assumed, both z_1 and z_2 are less than z_{up} due to the monotonicity of the subdifferential. Then, the inequality

$$\xi(f, \rho) \geq -\frac{z_1 + z_2}{P(S)} + \eta\ell(z_1) + (1 - \eta)\ell(z_2) \geq -\frac{2z_{\text{up}}}{P(S)}$$

holds for all $f, \rho \in \mathbb{R}$ and all η such that $\varepsilon \leq \eta \leq 1 - \varepsilon$. Hence, for any measurable function $f \in L_0$ and $\rho \in \mathbb{R}$, we have

$$\mathcal{R}(f, \rho) \geq \int_S \frac{-2z_{\text{up}}}{P(S)} P(dx) \geq -2z_{\text{up}}.$$

As a result, we have $\mathcal{R}^* \geq -2z_{\text{up}} > -\infty$.

B.2 Proof of Lemma 3

Corollary 5.29 of [27] ensures that the equality

$$\inf\{\mathbb{E}[\ell(\rho - yf(x))] : f \in \mathcal{H}\} = \inf\{\mathbb{E}[\ell(\rho - yf(x))] : f \in L_0\}$$

holds for any $\rho \in \mathbb{R}$. Thus, we have $\inf\{\mathcal{R}(f, \rho) : f \in \mathcal{H}\} = \inf\{\mathcal{R}(f, \rho) : f \in L_0\}$ for any $\rho \in \mathbb{R}$. Then, the equality

$$\inf\{\mathcal{R}(f, \rho) : f \in \mathcal{H}, \rho \in \mathbb{R}\} = \mathcal{R}^*$$

holds. Under Assumption 2 and Assumption 3, we have $\mathcal{R}^* > -\infty$ due to Lemma 2. Then, for any $\varepsilon > 0$, there exist $\lambda_\varepsilon > 0$, $f_\varepsilon \in \mathcal{H}$ and $\rho_\varepsilon \in \mathbb{R}$ such that $\|f_\varepsilon\|_{\mathcal{H}} \leq \lambda_\varepsilon$ and $\mathcal{R}(f_\varepsilon, \rho_\varepsilon) \leq \mathcal{R}^* + \varepsilon$ hold. For all $\lambda \geq \lambda_\varepsilon$ we have

$$\inf\{\mathcal{R}_\lambda(f, \rho) : f \in \mathcal{H}, \rho \in \mathbb{R}\} \leq \mathcal{R}_\lambda(f_\varepsilon, \rho_\varepsilon) = \mathcal{R}(f_\varepsilon, \rho_\varepsilon) \leq \mathcal{R}^* + \varepsilon.$$

On the other hand, it is clear that the inequality $\mathcal{R}^* \leq \inf\{\mathcal{R}_\lambda(f, \rho) : f \in \mathcal{H}, \rho \in \mathbb{R}\}$ holds. Hence, Eq.(30) holds.

B.3 Proof of Lemma 4

Under Assumption 2, the label probabilities, $P(y = +1)$ and $P(y = -1)$, are positive. We assume that the inequalities

$$\frac{1}{2}P(Y = +1) < \frac{m_p}{m_1}, \quad \frac{1}{2}P(Y = -1) < \frac{m_n}{m_1} \quad (41)$$

hold. Applying Chernoff bound, we see that there exists a positive constant $c > 0$ depending only on the marginal probability of the label such that (41) holds with the probability higher than $1 - e^{-cm_1}$.

Lemma 1 ensures that the problem (29) has optimal solutions $\hat{f}, \hat{b}, \hat{\rho}$. The first inequality in (31), i.e., $\|\hat{f}\|_{\mathcal{H}} \leq \lambda_{m_1}$, is clearly satisfied. Then, we have $\|\hat{f}\|_\infty \leq K\lambda_{m_1}$ from the reproducing property of the RKHSs. The definition of the estimator and the non-negativity of ℓ yield that

$$-2\hat{\rho} \leq -2\hat{\rho} + \frac{1}{m_1} \sum_{i=1}^{m_1} \ell(\hat{\rho} - y_i^{(1)}(\hat{f}(x_i^{(1)}) + \hat{b})) \leq \hat{\mathcal{R}}_{T_1, \lambda_{m_1}}(0, 0) = \ell(0).$$

Then, we have

$$\hat{\rho} \geq -\frac{\ell(0)}{2}. \quad (42)$$

Next, we consider the optimality condition of $\hat{\mathcal{R}}_{T_1, \lambda_{m_1}}$. According to the calculus of subdifferential introduced in Section 23 of [21], the derivative of the objective function with respect to ρ leads to an optimality condition,

$$0 \in -2 + \frac{1}{m_1} \sum_{i=1}^{m_1} \partial \ell(\hat{\rho} - y_i^{(1)}(\hat{f}(x_i^{(1)}) + \hat{b})).$$

The monotonicity and non-negativity of the subdifferential and the bound of $\|f\|_\infty$ lead to

$$\begin{aligned} 2 &\geq \frac{1}{m_1} \sum_{i=1}^{m_1} \partial \ell(\hat{\rho} - y_i^{(1)}\hat{b} - K\lambda_{m_1}) \\ &= \frac{1}{m_1} \sum_{i=1}^{m_p} \partial \ell(\hat{\rho} - \hat{b} - K\lambda_{m_1}) + \frac{1}{m_1} \sum_{j=1}^{m_n} \partial \ell(\hat{\rho} + \hat{b} - K\lambda_{m_1}) \\ &\geq \frac{1}{m_1} \sum_{i=1}^{m_p} \partial \ell(\hat{\rho} - \hat{b} - K\lambda_{m_1}). \end{aligned}$$

The above expression means that there exist numbers in the subdifferential such that the inequality holds, where $\sum_{i=1}^{m_p} \partial \ell$ denotes the m_p -fold sum of the set $\partial \ell$. Let z_p be a real number satisfying $\frac{2m_1}{m_p} < \partial \ell(z_p)$, i.e., all elements in $\partial \ell(z_p)$ are greater than $\frac{2m_1}{m_p}$. Then, $\hat{\rho} - \hat{b} - K\lambda_{m_1}$ should be less than z_p . In the same way, for z_n satisfying $\frac{2m_1}{m_n} < \partial \ell(z_n)$, we have $\hat{\rho} + \hat{b} - K\lambda_{m_1} < z_n$. The existence of z_p and z_n is guaranteed by Assumption 3. Hence, the inequalities

$$\begin{aligned}\hat{\rho} &\leq K\lambda_{m_1} + \max\{z_p, z_n\}, \\ |\hat{b}| &\leq \frac{\ell(0)}{2} + K\lambda_{m_1} + \max\{z_p, z_n\}\end{aligned}$$

hold, in which $\hat{\rho} \geq -\ell(0)/2$ is used in the second inequality. Define \bar{z} as a real number such that

$$\forall g \in \partial \ell(\bar{z}), \quad \max\left\{\frac{4}{P(Y=+1)}, \frac{4}{P(Y=-1)}\right\} < g.$$

Inequalities in (41) lead to

$$\max\left\{\frac{2m_1}{m_p}, \frac{2m_1}{m_n}\right\} < \max\left\{\frac{4}{P(Y=+1)}, \frac{4}{P(Y=-1)}\right\}.$$

Hence, we can choose \bar{z} satisfying $\max\{z_p, z_n\} < \bar{z}$. Suppose that $\ell(0)/2 \leq K\lambda_{m_1} + \bar{z}$ holds for $m_1 \geq M$. Then, the inequalities

$$|\hat{\rho}| \leq 2K\lambda_{m_1} + 2\bar{z}, \quad |\hat{b}| \leq 2K\lambda_{m_1} + 2\bar{z},$$

hold with the probability higher than $1 - e^{-cm_1}$ for $m_1 \geq M$. By choosing an appropriate positive constant $C > 0$, we obtain (31).

B.4 Proof of Lemma 5

Since $\|f\|_\infty \leq K\lambda_{m_1}$ holds for $f \in \mathcal{H}$ such that $\|f\|_{\mathcal{H}} \leq \lambda_{m_1}$, we have the following inequality

$$\begin{aligned}& \sup_{\substack{(x,y) \in \mathcal{X} \times \{+1,-1\} \\ (f,b,\rho) \in \mathcal{G}_{m_1}}} L(x,y;f,b,\rho) - \inf_{\substack{(x,y) \in \mathcal{X} \times \{+1,-1\} \\ (f,b,\rho) \in \mathcal{G}_{m_1}}} L(x,y;f,b,\rho) \\ & \leq 2C\lambda_{m_1} + \sup_{\substack{(x,y) \in \mathcal{X} \times \{+1,-1\} \\ (f,b,\rho) \in \mathcal{G}_{m_1}}} \ell(\rho - y(f(x) + b)) - (-2C\lambda_{m_1}) \\ & \leq 4C\lambda_{m_1} + \ell(C\lambda_{m_1} + K\lambda_{m_1} + C\lambda_{m_1}) \\ & = b_{m_1}.\end{aligned}$$

In the same way as the proof of Lemma 3.4 in [26], Hoeffding's inequality leads to the upper bound (35). Eq. (36) is the direct conclusion of (33) and (34).

C Proof of Theorem 2

Lemma 3 assures that, for any $\gamma > 0$, there exists sufficiently large M_1 such that

$$|\inf\{\mathcal{R}_{\lambda_{m_1}}(f + b, \rho) : f \in \mathcal{H}, b, \rho \in \mathbb{R}\} - \mathcal{R}^*| \leq \gamma$$

holds for all $m_1 \geq M_1$. Thus, there exist f_γ, b_γ and ρ_γ such that

$$|\mathcal{R}_{\lambda_{m_1}}(f_\gamma + b_\gamma, \rho_\gamma) - \mathcal{R}^*| \leq 2\gamma$$

and $\|f_\gamma\|_{\mathcal{H}} \leq \lambda_{m_1}$ hold for $m_1 \geq M_1$. Due to the law of large numbers, the inequality

$$|\widehat{\mathcal{R}}_{T_1}(f_\gamma + b_\gamma, \rho_\gamma) - \mathcal{R}(f_\gamma + b_\gamma, \rho_\gamma)| \leq \gamma$$

holds with high probability, say $1 - \delta_{m_1}$, for $m_1 \geq M_2$. The boundedness property in Lemma 4 leads to

$$P((\widehat{f}, \widehat{b}, \widehat{\rho}) \in \mathcal{G}_{m_1}) \geq 1 - e^{-cm_1}$$

for $m_1 \geq M_3$. In addition, by the uniform bound shown in Lemma 5, the inequality

$$\sup_{(f, b, \rho) \in \mathcal{G}_{m_1}} |\widehat{\mathcal{R}}_{T_1}(f + b, \rho) - \mathcal{R}(f + b, \rho)| \leq \gamma$$

holds with probability $1 - \delta'_{m_1}$. Hence, the probability such that the inequality

$$|\widehat{\mathcal{R}}_{T_1}(\widehat{f} + \widehat{b}, \widehat{\rho}) - \mathcal{R}(\widehat{f} + \widehat{b}, \widehat{\rho})| \leq \gamma$$

holds is higher than $1 - e^{-cm_1} - \delta'_{m_1}$ for $m_1 \geq M_3$. Let M_0 be $M_0 = \max\{M_1, M_2, M_3\}$. Then, for any $\gamma > 0$, the following inequalities hold with probability higher than $1 - e^{-cm_1} - \delta'_{m_1} - \delta_{m_1}$ for $m_1 \geq M_0$,

$$\begin{aligned} \mathcal{R}(\widehat{f} + \widehat{b}, \widehat{\rho}) &\leq \widehat{\mathcal{R}}_{T_1}(\widehat{f} + \widehat{b}, \widehat{\rho}) + \gamma \\ &\leq \widehat{\mathcal{R}}_{T_1}(f_\gamma + b_\gamma, \rho_\gamma) + \gamma \\ &\leq \mathcal{R}(f_\gamma + b_\gamma, \rho_\gamma) + 2\gamma \\ &= \mathcal{R}_{\lambda_{m_1}}(f_\gamma + b_\gamma, \rho_\gamma) + 2\gamma \\ &\leq \mathcal{R}^* + 4\gamma. \end{aligned} \tag{43}$$

The second inequality (43) above is given as

$$\widehat{\mathcal{R}}_{T_1}(\widehat{f} + \widehat{b}, \widehat{\rho}) = \widehat{\mathcal{R}}_{T_1, \lambda_{m_1}}(\widehat{f} + \widehat{b}, \widehat{\rho}) \leq \widehat{\mathcal{R}}_{T_1, \lambda_{m_1}}(f_\gamma + b_\gamma, \rho_\gamma) = \widehat{\mathcal{R}}_{T_1}(f_\gamma + b_\gamma, \rho_\gamma).$$

D Proof of Theorem 3

For a fixed ρ such that $\rho \geq -\ell(0)/2$, the loss function $\ell(\rho - z)$ is classification-calibrated [2], since $\ell'(\rho) > 0$ holds. Hence $\psi(\theta, \rho)$ in Assumption 4 satisfies $\psi(0, \rho) = 0$, $\psi(\theta, \rho) > 0$ for $0 < \theta \leq 1$, and $\psi(\theta, \rho)$ is continuous and strictly increasing in $\theta \in [0, 1]$. In addition, for all $f \in \mathcal{H}$ and $b \in \mathbb{R}$, the inequality

$$\psi(\mathcal{E}(f + b) - \mathcal{E}^*, \rho) \leq \mathbb{E}[\ell(\rho - y(f(x) + b))] - \inf_{f \in \mathcal{H}, b \in \mathbb{R}} \mathbb{E}[\ell(\rho - y(f(x) + b))]$$

holds. Details are presented in Theorem 1 and Theorem 2 of [2]. Here we used the equality

$$\inf\{\mathbb{E}[\ell(\rho - y(f(x) + b))] : f \in \mathcal{H}, b \in \mathbb{R}\} = \inf\{\mathbb{E}[\ell(\rho - y(f(x) + b))] : f \in L_0, b \in \mathbb{R}\},$$

which is shown in Corollary 5.29 of [27]. Hence, we have

$$\begin{aligned}\psi(\mathcal{E}(\hat{f} + \hat{b}) - \mathcal{E}^*, \hat{\rho}) &\leq \mathbb{E}[\ell(\hat{\rho} - y(\hat{f}(x) + \hat{b}))] - \inf_{f \in \mathcal{H}, b \in \mathbb{R}} \mathbb{E}[\ell(\hat{\rho} - y(f(x) + b))] \\ &= \mathcal{R}(\hat{f} + \hat{b}, \hat{\rho}) - \inf_{f \in \mathcal{H}, b \in \mathbb{R}} \mathcal{R}(f + b, \hat{\rho}),\end{aligned}$$

since $\hat{\rho} \geq -\ell(0)/2$ holds due to (42). We assumed that $\mathcal{R}(\hat{f} + \hat{b}, \hat{\rho})$ converges to \mathcal{R}^* in probability. Then, for any $\varepsilon > 0$, the inequality

$$\mathcal{R}^* \leq \inf_{f \in \mathcal{H}, b \in \mathbb{R}} \mathcal{R}(f + b, \hat{\rho}) \leq \mathcal{R}(\hat{f} + \hat{b}, \hat{\rho}) \leq \mathcal{R}^* + \varepsilon$$

holds with high probability for sufficiently large m_1 . Thus, $\psi(\mathcal{E}(\hat{f} + \hat{b}) - \mathcal{E}^*, \hat{\rho})$ converges to zero in probability. The inequality

$$0 \leq \tilde{\psi}(\mathcal{E}(\hat{f} + \hat{b}) - \mathcal{E}^*) \leq \psi(\mathcal{E}(\hat{f} + \hat{b}) - \mathcal{E}^*, \hat{\rho})$$

and the assumption on the function $\tilde{\psi}$ ensure that $\mathcal{E}(\hat{f} + \hat{b})$ converges to \mathcal{E}^* in probability, when m_1 tends to infinity. As a result, for any $\gamma > 0$,

$$|\mathcal{E}(\hat{f} + \hat{b}) - \mathcal{E}^*| \leq \gamma \quad (44)$$

holds with probability higher than $1 - \delta_{m_1, \gamma}$ with respect to the probability distribution of T_1 , where $\delta_{m_1, \gamma}$ satisfies $\lim_{m_1 \rightarrow \infty} \delta_{m_1, \gamma} = 0$ for any $\gamma > 0$.

Next, we study the relation between $\hat{f} + \hat{b}$ and $\hat{f} + \tilde{b}$. The sample size of T_2 is m_2 . For any fixed $f \in \mathcal{H}$, we define the set of 0-1 valued functions, $\mathcal{S}_f = \{\llbracket f(x) + b \geq 0 \rrbracket : b \in \mathbb{R}\}$. The VC-dimension of \mathcal{S}_f equals to one¹. Indeed, for two distinct points $x, x' \in \mathcal{X}$ such that $f(x) \geq f(x')$, the event such that $\llbracket f(x) + b \geq 0 \rrbracket = 0$ and $\llbracket f(x') + b \geq 0 \rrbracket = 1$ is impossible. Hence, for any $\varepsilon > 0$ and any $f \in \mathcal{H}$, the inequality

$$\sup_{b \in \mathbb{R}} |\hat{\mathcal{E}}_{T_2}(f + b) - \mathcal{E}(f + b)| \leq \gamma \quad (45)$$

holds with probability higher than $1 - \delta''_{m_2, \gamma}$ with respect to the joint probability of training sample T_2 . Note that $\delta''_{m_2, \gamma}$ depends only on m_2 , γ and the VC-dimension of \mathcal{S}_f . Thus, δ''_{m_2} is independent of the choice of $f \in \mathcal{H}$. Remember that $\hat{f} + \hat{b}$ depends only on the data set T_1 . Due to the law of large numbers, the inequality

$$|\hat{\mathcal{E}}_{T_2}(\hat{f} + \hat{b}) - \mathcal{E}(\hat{f} + \hat{b})| \leq \gamma$$

holds with probability higher than $1 - \delta'_{m_2, \gamma}$ with respect to the probability distribution of T_2 conditioned on T_1 . Since the 0-1 loss is bounded, it is possible to choose $\delta'_{m_2, \gamma}$ independent of \hat{f} . From the uniform convergence property (45), the following inequality also holds

$$|\hat{\mathcal{E}}_{T_2}(\hat{f} + \tilde{b}) - \mathcal{E}(\hat{f} + \tilde{b})| \leq \gamma$$

with probability higher than $1 - \delta''_{m_2, \gamma}$ with respect to the probability distribution of T_2 conditioned on the observation of T_1 . In addition, we have

$$\hat{\mathcal{E}}_{T_2}(\hat{f} + \tilde{b}) \leq \hat{\mathcal{E}}_{T_2}(\hat{f} + \hat{b}).$$

¹See [29] for the definition of the VC dimension.

Given the training samples T_1 satisfying (44), the inequalities

$$\mathcal{E}(\hat{f} + \tilde{b}) \leq \hat{\mathcal{E}}_{T_2}(\hat{f} + \tilde{b}) + \gamma \leq \hat{\mathcal{E}}_{T_2}(\hat{f} + \hat{b}) + \gamma \leq \mathcal{E}(\hat{f} + \hat{b}) + 2\gamma \leq \mathcal{E}^* + 3\gamma$$

hold with probability higher than $1 - \delta'_{m_2, \gamma} - \delta''_{m_2, \gamma}$ with respect to the probability distribution of T_2 conditioned on the observation of T_1 . Hence, as for the conditional probability, we have

$$P(\{T_2 : \mathcal{E}(\hat{f} + \tilde{b}) \leq \mathcal{E}^* + 3\gamma\} | T_1) \geq 1 - \delta'_{m_2, \gamma} - \delta''_{m_2, \gamma}.$$

Remember that $\delta'_{m_2, \gamma}$ and $\delta''_{m_2, \gamma}$ do not depend on T_1 . Hence, as for the joint probability of T_1 and T_2 , we have

$$P(\{T_1, T_2 : \mathcal{E}(\hat{f} + \tilde{b}) \leq \mathcal{E}^* + 3\gamma\}) \geq (1 - \delta'_{m_2, \gamma} - \delta''_{m_2, \gamma})(1 - \delta_{m_1, \gamma}).$$

The above inequality implies that $\mathcal{E}(\hat{f} + \tilde{b})$ converges to \mathcal{E}^* in probability, when m_1 and m_2 tend to infinity.

E Proofs of Lemma 6 and Lemma 7

E.1 Proof of Lemma 6

For $\theta = 0$ and $\theta = 1$, we can directly confirm that the lemma holds. In the following, we assume $0 < \theta < 1$ and $\rho \geq -\ell(0)/2$. We consider the following optimization problem involved in $\psi(\theta, \rho)$,

$$\inf_{z \in \mathbb{R}} \frac{1 + \theta}{2} \ell(\rho - z) + \frac{1 - \theta}{2} \ell(\rho + z). \quad (46)$$

The objective function is a finite-valued convex function on \mathbb{R} , and diverges to infinity when z tends to $\pm\infty$. Hence, there exists an optimal solution. Let $z^* \in \mathbb{R}$ be an optimal solution of (46). The optimality condition is given as

$$(1 + \theta)\ell'(\rho - z^*) - (1 - \theta)\ell'(\rho + z^*) = 0.$$

We assumed that both $1 + \theta$ and $1 - \theta$ are positive and that $\rho \geq -\ell(0)/2 > d$ holds. Hence, both $\ell'(\rho - z^*)$ and $\ell'(\rho + z^*)$ should not be zero. Indeed, if one of them is equal to zero, the other is also zero. Hence, we have $\rho - z^* \leq d$ and $\rho + z^* \leq d$. These inequalities contradict $\rho > d$. Then, we have $\rho - z^* > d$ and $\rho + z^* > d$, i.e., $|z^*| < \rho - d$. In addition, we have

$$\frac{1 + \theta}{2} = \frac{\ell'(\rho + z^*)}{\ell'(\rho + z^*) + \ell'(\rho - z^*)}.$$

Since $\ell''(z) > 0$ holds on (d, ∞) , the second derivative of the objective in (46) satisfies the positivity condition,

$$(1 + \theta)\ell''(\rho - z) + (1 - \theta)\ell''(\rho + z) > 0$$

for all z such that $\rho - z > d$ and $\rho + z > d$. Therefore, z^* is uniquely determined. For a fixed $\theta \in (0, 1)$, the optimal solution can be described as the function of ρ , i.e., $z^* = z(\rho)$. By

the implicit function theorem, $z(\rho)$ is continuously differentiable with respect to ρ . Then, the derivative of $\psi(\theta, \rho)$ is given as

$$\begin{aligned}
\frac{\partial}{\partial \rho} \psi(\theta, \rho) &= \frac{\partial}{\partial \rho} \left\{ \ell(\rho) - \frac{1+\theta}{2} \ell(\rho - z(\rho)) - \frac{1-\theta}{2} \ell(\rho + z(\rho)) \right\} \\
&= \ell'(\rho) - \frac{1+\theta}{2} \ell'(\rho - z(\rho)) \left(1 - \frac{\partial z}{\partial \rho} \right) - \frac{1-\theta}{2} \ell'(\rho + z(\rho)) \left(1 + \frac{\partial z}{\partial \rho} \right) \\
&= \ell'(\rho) - \frac{\ell'(\rho + z(\rho))}{\ell'(\rho + z(\rho)) + \ell'(\rho - z(\rho))} \ell'(\rho - z(\rho)) \left(1 - \frac{\partial z}{\partial \rho} \right) \\
&\quad - \frac{\ell'(\rho - z(\rho))}{\ell'(\rho + z(\rho)) + \ell'(\rho - z(\rho))} \ell'(\rho + z(\rho)) \left(1 + \frac{\partial z}{\partial \rho} \right) \\
&= \ell'(\rho) - \frac{2\ell'(\rho - z(\rho))\ell'(\rho + z(\rho))}{\ell'(\rho + z(\rho)) + \ell'(\rho - z(\rho))}.
\end{aligned}$$

The convexity of $1/\ell'(z)$ for $z > d$ leads to

$$0 < \frac{1}{\ell'(\rho)} \leq \frac{1}{2\ell'(\rho + z(\rho))} + \frac{1}{2\ell'(\rho - z(\rho))} = \frac{\ell'(\rho + z(\rho)) + \ell'(\rho - z(\rho))}{2\ell'(\rho - z(\rho))\ell'(\rho + z(\rho))}.$$

Hence, we have

$$\frac{\partial}{\partial \rho} \psi(\theta, \rho) \geq 0$$

for $\rho \geq -\ell(0)/2 > d$ and $0 < \theta < 1$. As a result, we see that $\psi(\theta, \rho)$ is non-decreasing as the function of ρ .

E.2 Proof of Lemma 7

We use the result of [2]. For a fixed ρ , the function $\xi(z, \rho)$ is continuous for $z \geq 0$, and the convexity of ℓ leads to the non-negativity of $\xi(z, \rho)$. Moreover, the convexity and the non-negativity of $\ell(z)$ lead to

$$\xi(z, \rho) \geq \frac{\ell(\rho + z) - \ell(\rho)}{z\ell'(\rho)} - \frac{\ell(\rho)}{z\ell'(\rho)} \geq 1 - \frac{\ell(\rho)}{z\ell'(\rho)}$$

for $z > 0$ and $\rho \geq -\ell(0)/2$, where $\ell(\rho)$ and $\ell'(\rho)$ are positive for $\rho > -\ell(0)/2$. The above inequality and the continuity of $\xi(\cdot, \rho)$ ensure that there exists z satisfying $\xi(z, \rho) = \theta$ for all θ such that $0 \leq \theta < 1$. We define the inverse function ξ_ρ^{-1} by

$$\xi_\rho^{-1}(\theta) = \inf\{z \geq 0 : \xi(z, \rho) = \theta\}$$

for $0 \leq \theta < 1$. For a fixed $\rho \geq -\ell(0)/2$, the loss function $\ell(\rho - z)$ is classification-calibrated [2]. Hence, Lemma 3 in [2] leads to the inequality

$$\psi(\theta, \rho) \geq \ell'(\rho) \frac{\theta}{2} \xi_\rho^{-1}\left(\frac{\theta}{2}\right),$$

for $0 \leq \theta < 1$. Define $\bar{\xi}^{-1}$ by

$$\bar{\xi}^{-1}(\theta) = \inf\{z \geq 0 : \bar{\xi}(z) = \theta\}.$$

From the definition of $\bar{\xi}(z)$, $\bar{\xi}^{-1}(\theta)$ is well-defined for all $\theta \in [0, 1]$. Since $\xi(z, \rho) \leq \bar{\xi}(z)$ holds, we have $\xi_{\rho}^{-1}(\theta/2) \geq \bar{\xi}^{-1}(\theta/2)$. In addition, $\ell'(\rho)$ is non-decreasing as the function of ρ . Thus, we have

$$\psi(\theta, \rho) \geq \ell'(-\ell(0)/2) \frac{\theta}{2} \bar{\xi}^{-1}\left(\frac{\theta}{2}\right)$$

for all $\rho \geq -\ell(0)/2$ and $0 \leq \theta < 1$. Then, we can choose

$$\tilde{\psi}(\theta) = \ell'(-\ell(0)/2) \frac{\theta}{2} \bar{\xi}^{-1}\left(\frac{\theta}{2}\right).$$

It is straightforward to confirm that the conditions of Assumption 4 are satisfied.

References

- [1] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *J. Comput. Syst. Sci.*, 54(2):317–331, 1997.
- [2] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.
- [3] A. Ben-Tal and A. Nemirovski. Robust optimization - methodology and applications. *Math. Program.*, 92(3):453–480, 2002.
- [4] A. Ben-Tal, L. El-Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, 2009.
- [5] K. P. Bennett and E. J. Breidensteiner. Duality and geometry in SVM classifiers. In *Proceedings of International Conference on Machine Learning*, pages 57–64, 2000.
- [6] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, 2004.
- [7] D. Bertsekas, A. Nedic, and A. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, Belmont, MA, 2003.
- [8] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.
- [9] D. J. Crisp and C. J. C. Burges. A geometric interpretation of ν -SVM classifiers. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 244–250. MIT Press, 2000.
- [10] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.
- [11] T. Evgeniou, M. Pontil, and T. Poggio. A unified framework for regularization networks and support vector machines. *Laboratory, Massachusetts Institute of Technology*, 1999.
- [12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, aug 1997.

- [13] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000, 1998.
- [14] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, New York, 2001.
- [15] G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2003.
- [16] D.G. Luenberger. *Optimization by vector space methods*. Series in decision and control. Wiley, 1997.
- [17] M. E. Mavroforakis and S. Theodoridis. A geometric approach to support vector machine (svm) classification. *IEEE Transactions on Neural Networks*, 17(3):671–682, 2006.
- [18] J. S. Nath and C. Bhattacharyya. Maximum margin classifiers with specified false positive and false negative error rates. In C. Apte, B. Liu, S. Parthasarathy, and D. Skillicorn, editors, *Proceedings of the seventh SIAM International Conference on Data mining*, pages 35–46. SIAM, 2007.
- [19] G. Rätsch, B. Schölkopf, A.J. Smola, S. Mika, T. Onoda, and K.-R. Müller. *Robust ensemble learning.*, pages 207–220. MIT Press, Cambridge, MA, 2000.
- [20] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine Learning*, 42(3):287–320, 2001.
- [21] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, USA, 1970.
- [22] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.
- [23] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
- [24] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.
- [25] I. Steinwart. On the optimal parameter choice for ν -support vector machines. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1274–1284, 2003.
- [26] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.
- [27] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008.
- [28] A. Takeda, H. Mitsugi, and T. Kanamori. A unified robust classification model, 2012. submitted.
- [29] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

- [30] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004.
- [31] D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.